# An investigation into the requirements of a big data stewardship training / instruction programme / curriculum

Mini-dissertation by

**Kyla Monique Yelverton**

Submitted in partial fulfilment of the requirements for the degree of

**MASTER OF INFORMATION TECHNOLOGY (B)**

in the

**FACULTY OF ENGINEERING, THE BUILT ENVIRONMENT AND INFORMATION TECHNOLOGY**

**UNIVERSITY OF PRETORIA**

Supervisors: Dr. MJ van Deventer & Prof T Bothma

*November 2019*

**Acknowledgements**

First and foremost, I would like to thank my heavenly Father for giving me the ability to be able to get to this stage of my academic career and to use my talents as well as possible.

Then I would also like to then extend my deepest gratitude to and express my appreciation for my supervisors, Dr. Martie Van Deventer and Professor Theo Bothma. I know the process was long, and sometimes very arduous and tedious, but I would never have been able to produce this mini-dissertation without their dedication, time, exceptional expertise and knowledge, and, of course, their constant support and motivation with regard to not only completing the mini-dissertation, but also to aspiring to deliver a high quality final product. I will forever remember the lessons learnt and the knowledge acquired from my supervisors.

Furthermore, I would like to thank everyone who played a significant role throughout this process, such as my parents, Noel and Joanne, who supported me financially and emotionally throughout my academic career. I would not be where I am today if it were not for their constant support in every possible way.

I would also like to thank my dearest Lionel, who supported me tirelessly through the most difficult, as well as the most pleasant times. Thank you for never letting me give up on this.

Last, but definitely not least, I would like to thank the University of Pretoria and the Department of Information Science for allowing me the opportunity to use their resources as well as to complete my master's degree at such a prestigious institution.

**Abstract**

Big data and their stewardship have become essential for environments such as the academia, as well as the corporate environment. With the ever-growing field of big data, it is becoming more and more important to be able to manage the data appropriately to achieve successful outcomes. This leads to the need for big data stewardship, which will assist with the appropriate curation of the big data at hand.

In order to ensure successful big data stewardship, the big data steward will need to fulfil specific roles and responsibilities, as well as have the necessary skills to fulfil those roles and responsibilities. Although different skills are necessary for different environments and situations, a general set of skills can be identified that are appropriate in different environments.

In order to adapt the skills needed for a big data steward, it was important to look at academic institutions to identify which programmes or courses these institutions are offering for the development of big data stewardship. These programmes indicate which skills are being addressed, as well as which skills are not being addressed but should be addressed.

The academic literature was also consulted to identify which roles and responsibilities big data stewards are expected to fulfil, as well as which skills the steward may need.

A thorough mapping analysis was done to identify the skills, roles and responsibilities needed for big data stewardship. This mapping assisted with the development of a recommended big data stewardship curriculum.

Accordingly, this research study aims to create a training benchmark so that big data stewardship can grow both in South Africa, and beyond.

# Contents

## List of Figures



## List of Tables

**Terms and definitions**

**Big data** - refers to "data sets that are so voluminous and complex that traditional data processing application software is inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, searching, sharing, transferring, visualising, querying, updating, information privacy and data sources. There are five dimensions to big data known as volume, variety, velocity and the recently added veracity and value" (Wikipedia, online) In turn, Press (2014) indicates that big data can be defined as data that are of a large size, typically to the extent where their manipulation and management present significant logistical challenges. Big data can thus be described as a large amount of data that need appropriate filtering and extraction to be of value.

**Long tail data** - The term refers to datasets that are usually, logistically speaking, less challenging to manage. (Genova & Horstmann, 2016: 6). Furthermore, long-tail data are the heterogeneous and smaller data which may have unique standards that are not regulated (Genova & Horstmann, 2016: 6). Long-tail data exist across all disciplines and are often without descriptive metadata. This creates a challenge for reusability. The characteristics of long-tail data will also differ with regard to their definition according to the size of the data, the format in which the data exists, the structure of the data, as well as the complexity of the data (Genova & Horstmann, 2016: 7).

**Business big data** -With regard to this type of data, business big data, it can be defined as a new form of capital for business in terms of it being a different and an innovative source of business value (Corea, 2016: 5). Furthermore, business big data "fulfils the role of optimising the scale, measurement, and calibration of big data, it models construction, validation, and visualization within the business, it assists in the creation of full datasets, it defines analytical frameworks, and it defines the business process more clearly" (Corea, 2016: 8). Business big data can thus be applied to any business or organisational environment and has the aim of improving the environment in which it exists.

**Big data stewards** - Big data stewardship, can be defined as the management and oversight of an organisations data assets in order to assist the organisation's users with high-quality and valuable data which is easily accessible and beneficial in use for the organisation (Rouse, 2013). Big data stewards thus have the main responsibility of ensuring that an organisation gains the most beneficial and valuable data that can be transferred into useful information

from big data sets, they are the intermediaries between big data and the use thereof for an organisation's workforce. For further reference within this research, data curation will also be used as a synonym for big data stewardship.

**Data life cycle** – The data life cycle is necessary for the successful management and preservation of data throughout its life cycle for use and reuse (DataONE, 2019). The data life cycle comprises of eight stages, these being namely to plan, collect, assure, describe, preserve, discover, integrate, and analyse data (DataONE, 2019). It is important to note that many versions of the data life cycle exist, but most of the versions have foundational stages, such as the planning stage. These data life cycles will differ in accordance with the expected outcome of the data throughout the life cycle, as well as the domain in which the data exists.

**Big data life cycle** - The big data life cycle comprises of all the steps and stages which big data needs to go through before it can be deemed useful and valuable to an organisation (Erl *et al.,* 2016). Erl et al. (2016), defines valuable stages of the big data life cycle namely data identification, data acquisition and filtering, data extraction, data validation and cleansing, data aggregation and representation, data analysis, data visualisation, and data utilisation of analysis results. One should keep in mind that the data steward is involved with the big data life cycle in order to ensure efficient curation of the data.

**Skill** – A skill can simply be defined by Merriam-Webster as "the ability to use one's knowledge effectively and readily in execution or performance." Within this research study, a skill is referred to as a data-driven skill which would include personal (soft) skills, and technical based skills. Therefore, a personal (soft) skill is defined as an interpersonal skill which may include communication skills, listening skills, and empathy skills (Doyle, 2019). Technical skills can be defined by Farley (2019) as "the knowledge and expertise needed to accomplish complex actions, tasks and processes relating to computational and physical technology."

**Competency** – Competencies are important in this research study as it goes hand in hand with skills, and the skill is needed to fulfil the competency. A competency can, therefore, be defined by the University of Nebraska-Lincoln (2019) as "a combination of observable and measurable knowledge, skills, abilities and personal attributes that contribute to enhanced performance by the individual for the given task." A competency in this research study is

needed as the steward needs to be competent to be able to curate the big data efficiently and successfully.

**Role** – A role needs to be defined for this research study as the steward needs to fulfil a specific role to be able to curate the big data successfully. A role can thus simply be defined by Merriam-Webster (2019) as "a role is a function or part performed by an individual in a particular process." The role, in this case, would be the data steward. The role of the data steward is defined to understand the responsibilities of the data steward.

**Responsibility** – A responsibility needs to be defined for this research study as the data steward needs to fulfil certain responsibilities which are identified in the literature review. A responsibility can, therefore, be defined by the Business Dictionary (2019) as "a duty which is conducted to achieve a specific task that one must fulfil, and which has a consequent success or failure." Furthermore, a responsibility can be defined as being responsible or accountable for certain actions to achieve a specific end goal.

**FAIR principles** – The FAIR principles are necessary to define for this research study as they play a large role in the literature review, as well as in the allocated and identified responsibilities of the data steward. The FAIR principles can thus be defined as data being Findable, Accessible, Interoperable, and Reusable (GoFair, 2019). The FAIR principles were created to provide guidelines to make data FAIR. The principles furthermore focus on machine learning because humans are relying more and more on computers to be able to handle and deal with the amount of data, which is constantly on the incline, thus being applicable for big data (GoFair, 2019)

# Chapter 1

## 1. Overview

### 1.1 Introduction

This research study was initiated because of a need to address big data stewardship in the context of an ever-growing dependency on big data – in research as well as in business. Hilbert (2016: 135) observes that an organisations' ability to cope with the uncertainty, which is caused by an ever-growing and developing economic, institutional and technological environment has become the main goal in the information age. Big data and their stewardship can present numerous techniques and solutions to address this uncertainty and can assist organisations by continuously helping them make informed decisions (Hilbert, 2016:136) and motivated by Kolb and Kolb (2013: 10).

Miemoukanda (2017) has thus identified the need and use for big data development and refers to the fact that organisations in South Africa have begun realising the potential of big data adoption and their benefits, and points out how these organisations should consider developing big data skills within their organisations.

This study will address and focus on the development of a big data stewardship curriculum for the University of Pretoria, South Africa and will consider relevant aspects of big data stewardship, such as the big data life cycle and its connection with the big data steward, as well as what can already be deduced from big data life cycle models and published literature for observation by big data stewards. The full understanding of big data stewardship will also be addressed as the challenges that big data stewards face will be considered, as well as how these relevant challenges will need to be overcome and addressed to perform big data stewardship.

In order to understand and comprehend all of the above, it is important to understand what the big data issues are, and why the interest in the stewardship of big data is so important. Pouchard (2015: 176) describes how researchers are using increasingly larger and more complex data to answer research questions. The capacity of the storage infrastructure, the increased sophistication and deployment of sensors, the abundant availability of computer clusters, the development of new analysis techniques, as well as larger collaborations, allow

researchers to address grand societal challenges in a way that is unprecedented because they have access to big data. Pouchard (2015: 176), thus illustrates how big data have become a part of a researcher's and a data analyst's everyday duties. However, little is known about what data stewards are doing to ensure that the data are reliable and remain accessible. Once again, this illustrates the importance of big data stewardship and its need.

Another aspect to keep in consideration throughout this study will be that of the five Vs of big data, namely, volume, velocity, variety, veracity and value (Castro, 2014: 15, Rahadi, Shobirin, & Ariyani, 2016: 45-47)). This aspect has been studied and addressed in this research, in terms of the big data life cycle and the curator's ensuing role. The intention is to show how the curator can make a difference in terms of the value of the output of big data (Castro, 2014: 15).

Furthermore, the research contributes to the knowledge pool of the community of big data stewards so that big data can be curated more effectively and efficiently. Accordingly, the data could prove to be an asset for any organisation's practices, whether the organisation is active in the private or public sector.

This researcher has shown how the data life cycle contributes to understanding the issues that need to be addressed when training curators to curate big data more effectively. This research can contribute to the identification of the role of the data steward within the big data life cycle, while also addressing the challenges, which the data steward may come across while curating big data sets. The research may also create a platform for further research to be done on big data stewardship and training on it with South African companies, and perhaps across national borders.

## 1.2 Objective(s)

The main objective of the research is to develop a well-researched curriculum for big data stewardship training that could be used by a prominent South African academic institution.

With the above-mentioned objective in mind the following sub-objectives were identified:

1.2.1. To gain a clear understanding of the framework and/or context for big data.

1.2.2. To define the role that data stewards could play within a generic big data life cycle.

1.2.3. To document the known challenges of big data stewardship.

1.2.4. To establish how institutions of higher learning (international as well as South African universities) are approaching the challenge to train big data stewards.

These objectives were then unpacked into a number of research questions and sub-questions.

## 1.3 Central research question and sub-questions

The next section discusses the main research question as well as the sub-questions. These questions will guide the study as to what needs to be solved and what are the key issues which are being analysed.

**Central research question**

What would be a realistic curriculum, for training big data stewards, at a South African university?

**Formulation of sub-questions or sub-problems**

Several further questions were formulated to underpin the central question. These are:

1.3.1 What is the current framework or context for big data? (Refer to section 2.1)

- What are the key concepts that a big data stewards should be familiar with? (Refer to section 2.3 and 2.4)
- How does big data stewardship differ from long tail data stewardship? (Refer to section 2.4.2)
- When is big data an asset? (Refer to section 2.4.4)

1.3.2 How does data stewardship fit into the big data life cycle? (Refer to section 2.2)

- Which big data life cycles are being promoted? (Refer to section 2.6)
- What are the characteristics of the big data life cycle models? (Refer to section 2.6.1)
- How do the cycles used in the big business data differ from the cycle models used for big research data? (Refer to section 2.6.3)
- Which stewardship activities could be linked to a preferred big data life cycle model? (Refer to section 2.6)

1.3.3 What are the known challenges of big data stewardship training? (Refer to section 2.3)

- What is the desired skill set for big data stewards? (Refer to section 2.9.11)
- Which ownership responsibilities do data stewards have in the big data life cycle? (Refer to sections 2.5 – 2.7)
- Which courses and training are available for big data stewards? (Refer to Chapter 4)

1.3.4 How is a selected group of universities (international and South African) approaching the training of big data stewards? (Refer to Chapter 5)

1.3.5 With regard to the online content - how is big data stewardship education being addressed by South African university departments and Library and Information Science schools?

## 1.4 Scope and limitations

The scope of the study entailed the target group at which this research was aimed, the geographical location where the physical research took place, and the time period it would take for the study to be completed.

The target group of this research included tertiary institutions within South Africa, as well as tertiary institutions on the World University Rankings by Times Higher Education ("World University Rankings", 2018).

The place where the research took place was at the University of Pretoria, as the research utilised observations as well as content analysis.

This study was conducted throughout 2018, while the data analysis took place in 2019. During this period ethics clearance was also obtained. It is anticipated that the period was long enough to gather the data after which data analysis and interpretation took place.

As far as the limitations of the research are concerned: the first limitation would be that the research was limited to the University of Pretoria. Further studies will be required to extrapolate the results to big data stewardship at national level.

Another limitation is that not all disciplines within the university were used for this research. Only specific departments within the university were preselected.

Another limitation was that this field of study is extremely dynamic, and it is acknowledged that new literature was published after 3 April 2019, the date all the literature was accessed for the purpose of this study. The researcher could not monitor the literature and sources which were published after the above date, which is also an acknowledgement that new literature regarding the topic could have been published. Similarly, the content on all of the websites consulted could have also changed throughout the duration of the study.

The last limitation would be that of time. An extremely limited period of time was available to conduct this research. This can, however, be overcome by good time management with regard to this study.

### 1.5 Justification for the research

The reason or justification for this research is that the need for trained data stewards – especially big data stewards is due to increase in an extremely short period of time (Pouchard, 2015: 176). South Africa will be experiencing the same shortage of skilled data stewards. The purpose of this research is to identify the specific training programme(s) needed to train big data stewards so that they can have a larger and more positive impact on institutions in South Africa – in both the private and public sectors.

From the published literature, big data stewardship and the training programmes for stewards relate to a relatively new field and subject matter not yet fully explored as a research subject both within South African universities as well as South African businesses (Russom, 2017: 2). This research can assist with adding to the body of knowledge available to data stewards.

Furthermore, an attempt has been made to identify a definitive big data life cycle in order to analyse the role of the big data stewards in the given life cycle. The evaluation of the life cycle and its related aspects will also add value to the subject matter and can perhaps assist with identifying the role of the steward within the life cycle. This can also add value to future research and training for big data stewardship.

Furthermore, this research may act as a contributor to fields such as knowledge management and information science within the UP Department of Information Science, thereby contributing to high-quality training provided to future scholars.

Furthermore, this research can act as a base on which to build and add onto. Big data (and the stewardship thereof) is a constantly growing field. Providing stewardship services can assist in maturing the use of big data in South Africa.

### 1.6 Overview of the literature

The literature that was applicable for this research can be subdivided into three different categories, namely, (1) the curation and management of big data (research and business), (2) the skills, competencies and roles of data stewards (3) general information to use as a foundation to provide context to research on big data stewardship.

Different perspectives on big data curation, such as scale and complexity, policy, and value of management, and the library's role in curating and exposing big data contributed towards this research. The literature (for example, Pryor, 2012: 2; Teets & Goldner, 2013: 429-438; Castro (2014: p15-18; Pouchard, 2015: 176-190) provided an understanding of big data curation from a variety of different perspectives. In turn, this assisted in building an understanding of big data stewardship – which, in the end, contributed to an understanding of the big data steward and the role of the steward in the big data life cycle.

Another article which was useful for this research was an article that by Rossi and Hirama (2015: 165-180) that focusses on the characterising of big data management. This article looks at the challenges that organisations may face regarding big data, such as big data modelling, storage and retrieval, analysis and visualisation, as well as the consideration of how people are necessary for the facilitating of big data management (Rossi & Hirama, 2015: 165-180).

Rahadi, Shobirin and Ariyani, (2016: 45-47) describe big data management in terms of the five V's of big data which are volume, velocity, value, veracity, and variety. Furthermore, the article considers how the term "big data" applies to large volumes of data that can either be structured or unstructured, and which overwhelms an organisation on a daily basis (Rahadi *et al.,* 2016: 45-47). The article thus considered the concepts, tools and technology used for big data, and a big data-based product ranking solution which can be considered by

organisations, this article thus intends on achieving simpler big data management within all types of organisations (Rahadi *et al.,* 2016: 45-47).

Articles which relate to big data for business were obtained from authors such as Bughin (2016: 1-4, Russom (2017:1-8), Corea (2016: 5-17), as well as Saltz (2015:2872-2879), all of whom focussed on big data and getting a better performance from big data, and who also note that organisations have made significant investments in data warehouses and analytical programmes to benefit from and derive value from big data management within the organisation. These articles further identify the returns on big data investments and identify data analytics as well as their profitability and value-added productivity. In addition, they also discuss the modern integration and data quality practices for digital business requirements. Furthermore, these authors mention identify a need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness.

In order to understand and grasp the concept of 'big data' more fully, articles which discuss the background and development of big data were also used to add value to this research. These articles are by authors such as Cai and Zhu (2015: 1-10) and Cockayne (2016: 1-11). These authors evaluate the challenges of data quality and data quality assessment in the big data era, thereby providing a broader understanding of big data in any environment and any organisation. The articles by the authors mentioned above, include the concepts of 'affect' and 'value' in critical examination of the production and the presumption of big data is considered, which explores the relationship between production and the value of big data in any organisation or environment.

Articles by Zwitter (2014: 1-6), as well as Lagoze (2014: 1-11) were used as they deal with big data ethics, which is an integral part of any big data research as it lays the foundation of big data and its eligible functions Furthermore, the articles include important content on big data, big data integrity, and the fracturing of the control zone of big data, which evaluate the paradigm shift of big data in today's every day big data affiliations and tasks.

Loukissas (2016: 1-20), and Coates (2014: 52-59) have written articles that include the place of big data in the organisation; while close distant readings of accessions data from the Arnold Arboretum are mentioned. Importantly, the best place for big data to be used beneficially and valuably in any given environment is also touched on. In addition, they refer to the building

of data services from the ground up, including the big data strategies and big data resources to be used.

The articles mentioned above, served as the starting point for the literature survey. Further articles and books were identified and studied to answer and understand the main and sub-objectives of this study. With the evaluation of the above articles, a further link was made to what the current state of big data stewardship is and what it should entail when efficient and valuable training is in place.

## 1.7 Research methodology

The research methodology section discusses the research paradigm, the research design, the data collection methods, the data collection tools, the target population, as well as the sampling method.

The **research methodology,** which was used for this study is qualitative. The reason for this is that the data which were collected were detailed and provided an in-depth understanding of the issues that confront an extremely specific group of people – big data stewards. According to Atieno (2009: 16) and further motivated by Crossman (2017), there were several advantages or strengths of using a qualitative research method, one being that qualitative research is able to manage and clarify the data without spoiling its complexity and context.

Furthermore, qualitative research can be described as a study method, which allows the researcher to make sense of and to understand complex situations better, meaning that observations are developed from the ground in order to build an accurate theory of the situation (Leedy & Ormrod, 2015: 98). Qualitative research can be associated with case studies to gain valid and in-depth data, whereas quantitative research is concerned with generalisations of the data (Leedy & Ormrod, 2015: 99). Qualitative research provides more in-depth and detailed data, it creates openness between the researcher and the participant as the researcher could make use of research tools, which allow open-ended questions and discussions to be held with the participant to gain more data regarding the research. This also means that the researcher has to enter the study with an open mind to interact personally with the participants of the research and to gain the most detailed and perhaps holistic data possible (Leedy & Ormrod, 2015: 99).

Furthermore, qualitative research simulates people's individual experiences, and it attempts the avoidance of pre-judgements and bias as it provides an explanation of the specific data captured and why the data are as they are (Atieno, 2009: 17) and supported by Lewis (2015: 473-475). The disadvantages of qualitative research, as discussed by Datt and Datt (2016), is that the process is time-consuming and may take much more time for data analysis than is the case with quantitative research. Secondly, the data collected by the participant cannot be verified in terms of making sure that the data are not only based on the participant's perception but is also objective. Qualitative research entails a labour-intensive approach as the researcher has to analyse and derive conclusions from the more than likely wide variety of data, which have been captured, and, lastly, the fact that qualitative research is difficult to use to investigate causality, meaning that researchers may find it difficult to establish which event led to which reaction. The research method, however, is largely influenced by the researcher's thinking processes, and often the researcher will choose a method that suits their analysis and thinking skills.

**Data collection instruments** were the internet, websites, as well as relevant internet sources from the selected tertiary institutions regarding programmes on big data management. The data collection instrument is illustrated in the form of a template, which illustrates all the data collected.

The data collected were analysed by means of content analysis that can be defined as a detailed and systematic examination of specific contents from specific material for the purpose of identifying patterns, themes, or biases within the analysed material (Leedy & Ormrod, 2015: 102). Content analysis was used for this study as it is a type of analysis, which involves the analysis of specific content such as websites, internet sources, and other applicable sources. The purpose of the content analysis was to make informative judgements and recommendations based on the analysis.

### 1.8 Target population and sampling

The main target population for this research was tertiary institutions in South Africa, as well as selected universities from the top 100 ranked universities internationally. The sampling technique used for this study was systematic sampling as the participants within the target

population were electedsystematically s because of the value which they gave to the research study.

The reason for choosing this sample population is because of the valuable information and insights that can be gained from the participants regarding the validity of the training programme identified in the literature.

### 1.9 Value of the study

This study and its research can be deemed to be valuable as it can contribute further to research based on big data stewardship, which is necessary in South Africa as there was not much evidence that similar research has already been conducted in South Africa. This research study is also valuable as it can expand on the existing knowledge regarding big data stewardship in South Africa, as well as internationally.

### 1.10 Ethical clearance

As the information, which was considered, as well as the data which were collected exist in the public domain, it was not necessary to obtain ethical clearance from the University of Pretoria. It is, however, necessary for a researcher to apply due diligence throughout the research process to ensure that the information and data that were collected, were treated with the necessary respect and in an appropriate manner.

### 1.11 Division of chapters

Chapter 1 provides the objectives for the research and includes the justification, the value, the scope and the limitations of the study. It also provides the methodology which was used to conduct the research and indicates how the data were analysed.

Chapter 2 includes an analysis of the available literature regarding the topic. The literature review aimed to enhance the richness of the research and its purpose and anticipated how some of the questions guiding this study would be answered.

Chapter 3 takes the literature review in Chapter 2 further. Chapter 3 includes the identified literature regarding the roles and responsibilities of the data steward, as well as the skills identified with regard to the data steward.

Chapter 4 includes a discussion of the chosen research methodology as well as why the specific methodology was chosen. This chapter will also include an appendix with the data collection tool that will be used to complete the research.

Chapter 5 includes the analysis of the data collected by means of the data collection instrument. This chapter then includes how the data was analysed and how the analysed data can be used to answer the main and sub-questions of the research.

Chapter 6 concludes the study and summarises answers to the questions indicated as the purpose of the research. It introduces recommendations for future studies regarding the topic.

## 1.12 Conclusion

In conclusion, this chapter has addressed the problem statement as well as the sub-problem statements as well as the objectives of the study to help provide clarification with regard to big data stewardship at the University of Pretoria and identify the training needed for big data stewardship.

Big data may seem easy to collect, but the process that the data have to undergo to make the data valuable, is difficult and complex. From this, one can recognise the importance of appropriate training for big data stewards. This research is thus aimed at achieving the identification of appropriate content for a data stewardship training programme/curriculum so that the stewards can perform their roles successfully and efficiently and, in the end, produce data products that are FAIR – findable, accessible, interoperable and reliable, for all who have contributed financially to the collection of the research data or for those who are able to take the research further.

# Chapter 2

## 2. Literature review

### 2.1 Introduction

With big data developing rapidly, and with an awareness of it growing in the private and the public sectors, it is becoming crucial for all organisations to have the capability to manage and curate big data properly. Big data can play a crucial role in the future business of any organisation, but the organisation must first understand the qualities of big data before their potential contribution to a paradigm shift can be appreciated (Lagoze, 2014: 1). With these factors in mind, organisations, and more specifically, those members of staff responsible for managing knowledge assets, should begin focussing on the big data life cycle. Of further importance is the curator's role in the big data life cycle: managing big data efficiently and, where appropriate, packaging it effectively when the end of the current life cycle is reached. Properly curated data can lead to data products that will extend the useful life of the data. It is expected that, with the addition of big data skills, many organisations will learn valuable facts regarding their processes, operations, customers and every other related facet of the organisation (Russom, 2013: 4).

The question that comes to mind is: What training and expertise are needed for big data curators so that they can play their roles successfully? This research focusses on big data from the perspective of an academic institution wishing to educate curators. The author considers what training big data curators should receive to have a positive impact on the work environment. The big data curator training that is currently available internationally and, in South Africa, was investigated. The result of that investigation is reported in Chapter 4.

Both research and business big data are studied in this research because big data curation is applicable to both private and public-sector organisations. Big data management also has several benefits for the organisation – be it for research or commercial purposes (Russom, 2013: 4). This study should be insightful and beneficial for all stakeholders as the content could be a foundation for big data curation growth and awareness in all the sectors.

## 2.2 Big data defined

To evaluate, understand and use' big data,' this concept must first be understood fully and defined.

Big data can be defined as extremely large datasets that include masses of unstructured data that must be analysed for further use (Chen, Mao, & Liu, 2014: 171). These masses of data cannot be captured, managed, and processed by traditional information technology (IT) hardware and software tools and technology within an acceptable period of time and scope (Chen *et al.,* 2014: 173).

Furthermore, big data refer to large and diverse amounts of data that cannot be processed in a simple manner. Big data are usually a set of data existing within a database with a volume that exceeds the typical size of a normal database (Rossi & Hirama, 2015: 165). Big data require "new" technology to handle their creation, storage, management, and analysis (Rossi & Hirama, 2015: 165). Big data thus entail the conception of a large amount of data that cannot be handled efficiently, hence, requiring new tools and technology to handle, process and analyse it successfully and efficiently (Rossi & Hirama, 2015: 165). Big data are better described through their characteristics, by comparing them to long tail data and by evaluating the importance of their value as both information and knowledge assets.

### 2.2.1 Characteristics of big data

Big data can also be understood by looking at the characteristics of the data. These characteristics are often expressed as the 'Vs' of big data. The number of 'Vs' keep on expanding – starting first with three Vs (volume, variety and velocity) to a point where 42Vs are currently considered (Shafer, 2017). For this study, only the five best known characteristics (5V-concepts) will be used. By this is meant understanding big data in terms of their volume (or size), the variety of data which relates to the large number of variables, the velocity of data, which includes the speed with which streams of data are generated, exist and are processed, the veracity of the data that includes the quality and relevance of the data, and, lastly, the value of the data (Rahadi, Shobirin, & Ariyani, 2016: 45). When considering the 5V characteristics, it is already possible to understand big data better and their associated qualities. The 5Vs also assist organisations to prepare better for the impact of their big data,

as it is not only the volume of data which is important; what the organisation does with the data is what matters most. For the data steward it is especially the veracity and the value of the data that are of importance.

As the first V is the volume of big data; it is important to note that the amount of data that exists in the world is doubling every two years (ISO/IEC JTC 1, 2014: p9). Organisations now work with requirements for analytical data volumes in terms of terabytes and petabytes (ISO/IEC JTC 1, 2014: 9; Rahadi *et al.* 2016: 45).

The second V is variety, which refers to whether the aspects of the big data are structured or unstructured, or if the data exist in relational formats or not (Rahadi *et al*., 2016: p45). Furthermore, variety includes diverse application domains in which the big data exist, where applications create, consume, process and analyse the data at hand (ISO/IEC JTC, 2014: p10). In a social media setting big data will be defined as large and will involve diverse amounts of unstructured data, such as tweets, videos, images, and audio clips (Pouchard, 2015: 179).

The third V is velocity, which includes the rate at which the data are created, stored, analysed, and visualised. Data flow rates are increasing considerably in speed and variability, which create the challenges of enabling real or near real-time data usage (ISO/IEC JTC, 2014: p10).

The fourth V is veracity, which includes the trustworthiness of the data, the authenticity of the data, the origin of the data, the availability of the data, and the accountability of the data (Rahadi *et al.,* 2016:45). The veracity of big data is essential to the value of big data, as the data are applicable to a specific problem or challenge (ISO/IEC JTC, 2014: p11).

The fifth V is the value of big data. This characteristic is concerned with the statistical value of the data, the correlations of the data, and the hypothetical value of the data (Rahadi *et al.,* 2016: 45). ISO/IEC JTC (2014: 10) discusses the value component of big data in terms of variability, which includes the changes that take place in the data's rate, format, structure, semantics, and quality, which, in the end, influence the supported application of the big data. The impact of the value perception refers to the big data architecture, interface, algorithms, integration, storage, applicability, and use (ISO/IEC JTC, 2014: 11).

To ensure that big data are of the appropriate standard within any given organisation and

environment, it is important to note the veracity of big data, as this is the characteristic that ensures that big data remain valuable. As stated by Buhl, Roglinger, Moser& Heidemann (2016: 67), the meaningful use of data veracity is underpinned by a clear data governance and data policy. This clear governance and data policy are only enabled when high data quality exists, and in order for high data quality to exist, it needs to entail consistency, content, meaning, completeness, comprehensibility, reliability, as well as data allowing for unique identifiability (Buhl *et al.,* 2013: p66-67).

Rossi (2015) identifies the value of big data in terms of volume, velocity, variety, and veracity, as the four other characteristics of the five Vs. Volume-based value includes the ability of organisations to store larger volumes of data, which means that the sampling of data is no longer necessary as organisations can analyse the data as one piece (Rossi, 2015). Organisations storing larger volumes of data may lead to better decision-making regarding the acquiring, increasing and managing of data for operations (Rossi, 2015).

The veracity-based value of big data includes the speed at which organisations can manage their data, including the injection of data into data platforms (Rossi, 2015). The rapid analysis of big data, made possible by the increased speed of big data operations, assists organisations to make faster and better decisions regarding their objectives and operations (Rossi, 2015) and further stated by Buhl *et al.* (2013: 68).

Variety-based value involves the organisation being able to acquire and analyse a variety of data regarding customers and operations (Rossi, 2015). The greater variety of big data the organisation has, the more multi-faceted the organisation can be in fulfilling needs within the organisation (Rossi, 2015). Variety-based value creates deeper insights into the existing datasets, that can assist with the development and personalisation of data for different purposes (Rossi, 2015).

The veracity-based value of big data includes the quality of the data within the big data context, and the organisational propositions, which can be developed from the large amount of data in the organisation (Rossi, 2015). The accuracy of the data within the organisation is also important but is not as crucial at the beginning stages of the design and validation of the data (Rossi, 2015).

Furthermore, to add to the discussion of the extremely important 'Vs,' which assist in defining big data, Fosso *et al*. (2015: 5-6) recognise the 5 Vs and explain the different natures of each V. Wamba, Akter, Edwards, Chopin & Gnanzou. (2015: p5-6) explains how the first V, which is volume, is a large amount of data which has the function of either consuming huge storage or consists of a large amount of records, as supported by Russom (2011). An example of this within an organisation would be data warehousing, which includes a large number of petabytes of information (Manyika *et al.,* 2011).

The second V discussed by Wamna *et al.* (2015: 5-6), and further supported by Russom (2011), is that of variety, where variety's nature is explained as data generated from different sources and platforms and, therefore, contains multidimensional data formats and fields. An example of this within an organisation would be a platform on which each department works and shares data, meaning that each department within the organisation is different and contains different forms of data, which all come together in the end and form multidimensional data fields on the same platform.

The third V discussed by Wamba *et al*. (2015: 5-6) that is further elaborated on by Russom (2011), is that of velocity. The nature of velocity can be discussed as the regularity of data being generated, or the regularity of the data being delivered (Russom, 2011). An example of this within an organisation would be if a delivery company constantly took new orders for products, but was not able to provide delivery dates for those products, meaning that the frequency of data generation for the system may not compromise the frequency of data delivery (Davenport, 2006).

The fourth V discussed by Wamba *et al.* (2015: 5-6) is that of veracity that can be described in terms of its nature as the data is unpredictable, which in the end, requires the proper analysis of big data to gain a reliable and efficient prediction of the actual initial data within the system (Beulke, 2011). An example of this would be the replication of data on the same platform, causing an interference with the system to predict data accurately. This would mean that an internal system would need to be created in order to filter the data accurately which would minimise data replication and inaccuracy (Davenport, Barth & Bean, 2012).

The fifth and final V as discussed by Wamba *et al*. (2015:5-6) is that of value. The nature of value lies in the extent to which the big data provide worthy and valuable benefits and

understanding within their context through the process of extraction and transformation into a final and valuable product (Wamba *et al*., 2015: 5-6). An example of this would be any organisation using their data efficiently to predict customer wants and needs, to enhance and improve organisational procedures and processes.

From all that has been discussed above, regarding the Vs of big data, it is important to note that the value and veracity are more important for the curator/steward, which is applicable to the study. As the steward is involved with the classification and management of the data, it is important to link value and veracity to the steward's duties. Firstly, the data steward needs to add value to the data, which means adding more details or any data elements which will make the data highly valuable. This means that the steward needs to keep the context in mind in which the data exist, as well as the role the data are supposed to play as an end product (Saagie, 2017). By adding value to the data, the steward assists the user of the data to make informed decisions which will make the data more useful and valuable than what it would have initially been if the data steward did not go through the process of adding value (Biehn, 2018).

The veracity of data is also closely linked to the role of the data steward as the data steward needs to ensure that the data are in the correct form for proper analysis. It is also the duty of the steward to analyse the data successfully to ensure that the veracity and relevance of the data is not doubted (Saagie, 2017). As the data steward is involved in producing the end product of data to the user of the data, it is important that the steward take note of the unpredictability of the data so that they can analyse the data to ensure their reliability and quality to add value for the user of the data (Biehn, 2018).

Teets and Goldner (2013: 431) state that:

> To search successfully for new science in large datasets, we must find the unexpected patterns and interpret evidence in ways that frame new questions and suggest further explorations. Old habits of representing data can fail to meet these challenges, preventing us from reaching beyond the familiar questions and answers.

Big data entail an ever-growing concept, which present many new opportunities that can be beneficial for the organisation, depending on how the data are managed and used.

### 2.2.2 Difference between big data stewardship and long tail data stewardship

To understand the difference between big data stewardship and long tail data stewardship, one should first understand the difference between big data and long tail data. Where big data are defined as very large datasets, which include large amounts of unstructured data that must be analysed for further use (Chen, Mao & Liu, 2014: 171), long tail data can be defined as heterogeneous, and relatively small in file-size data, which have unique standards, which are regulated, requiring personalised curation and control within smaller domain repositories in which the long tail data may exist (Genova & Horstmann, 2016: 6). It is often assumed that long tail data are only the remaining data, or the less important parts of the data. This denigrates long tail data and is not a true definition (Genova & Horstmann, 2016: 6).

Big data therefore focus on the exponential growth of data generation and availability, consisting of structured and unstructured data, while long tail data focus on the variety in structure, subject, complexity, format, size, location, and use of the data in the context of research (Heidorn, 2008), with which Boyd and Crawford (2012), as well as Borgman (2015) concur.

According to Heidorn (2008), there are specific characteristics of ong tail and big data which assist in understanding the role of each, as well as the nature and purpose of each type of data. Big data for example are homogenous, while long tail data are heterogenous, big data are large, while long tail data are smaller in size (Heidorn, 2008). Big data must conform with more common standards, while long tail data must conform with unique standards; furthermore, big data are regulated, while long tail data are not regulated (Heidorn, 2008). Big data exist in disciplinary repositories, while long tail data exist in institutional, general, or no repository at all (Heidorn, 2008). Lastly, and most important for the focus of this study, big data are characterised by and need central curation, while long tail data need individual curation (Heidorn, 2008).

Because of the sheer volume, it is more likely for big data to be centrally curated at the point of collection, while long tail data is usually curated where it makes most sense to do so. Big data are usually not accessible via a repository, while long tail data are found in either

discipline or institutional repositories (Genova & Horstmann, 2016: 7). It is anticipated therefore, that the stewardship skills required, may also differ.

### 2.2.3 Big data as the foundation of valuable information – a format to be managed

One can deduce from Lagoze (2014: 1) that big data have become an asset in the organisation, public or private, as big data simply exist in today's world and they cannot be ignored. For big data to be regarded as an asset by the organisation, they must have economic, research and/or social value. The data should be in a usable form(at) so that they can become actionable information (Cockayne, 2016: 9). This also means that for the data to be of value, they must be managed and curated properly (Cockayne, 2016: 9). The individual(s) responsible for managing the big data should know what value the big data can produce for the organisation (Cockayne, 2016: 9). When considering the value, the proper curation of big data can produce, the training and expertise of the steward do need special attention.

There are three points of view to consider in big data management: big data collectors/ generators, big data utilisers, and big data curators. The big data collectors determine which data are to be collected and how long they should be stored (Zwitter, 2014: 3). Big data generators include natural actors of big data, artificial actors of big data, and physical phenomena which also generate large amounts of data in their natural existence (Zwitter, 2014: 3). Big data utilisers, who are part of the utility side of production of the big data, also redefine the purpose of the collection of the data within the big data management process of the organisation (Zwitter, 2014: 3). Big data curators need to ensure that the data generated or collected remain accessible to the data utilisers long after the collection process has been completed.

### 2.2.4 Big data as a knowledge asset – to be stewarded

Lagoze (2014: 1) indicates how the business and public sectors both show enthusiasm regarding big data. For business, big data present new possibilities for direct and micro marketing, supply-chain optimisation, and other ways of producing or increasing efficiency and profit. In the public sector, big data offer new possibilities and opportunities in areas such as security and terrorism prevention (Lagoze, 2014: 1).

As much as big data are an asset when they become information and they have economic and social value for the organisation, they may also have value for the organisation in the form of knowledge. Cawthorne (2015) discusses how big data can be analysed and transferred into a knowledge asset for the organisation. Big data are analysed with the aid of information models and trends by data scientists. The data steward then captures and curates the methodologies which were used to analyse the data (Cawthorne, 2015). This then becomes business intelligence, which is valuable to the organisation, as it shows what happened to actionable data in the past and how it succeeded or failed, and what the analysed and interpreted data imply, should be done differently to ensure success (Cawthorne, 2015). This is especially true for business intelligence when the data become actionable insights through correct and efficient analysis and interpretation of the big data for employees to use.

### 2.2.5. Big data and knowledge management

Although it is possible to make several assumptions that have relevance for big data as knowledge assets, extremely little has been documented regarding the relationship between big data and knowledge management. A study conducted in Nigeria, a developing country, was the only one retrieved that makes a direct link between knowledge management and big data. Alhinn and Rababah (2018: 6) illustrate this relationship in the figure below.

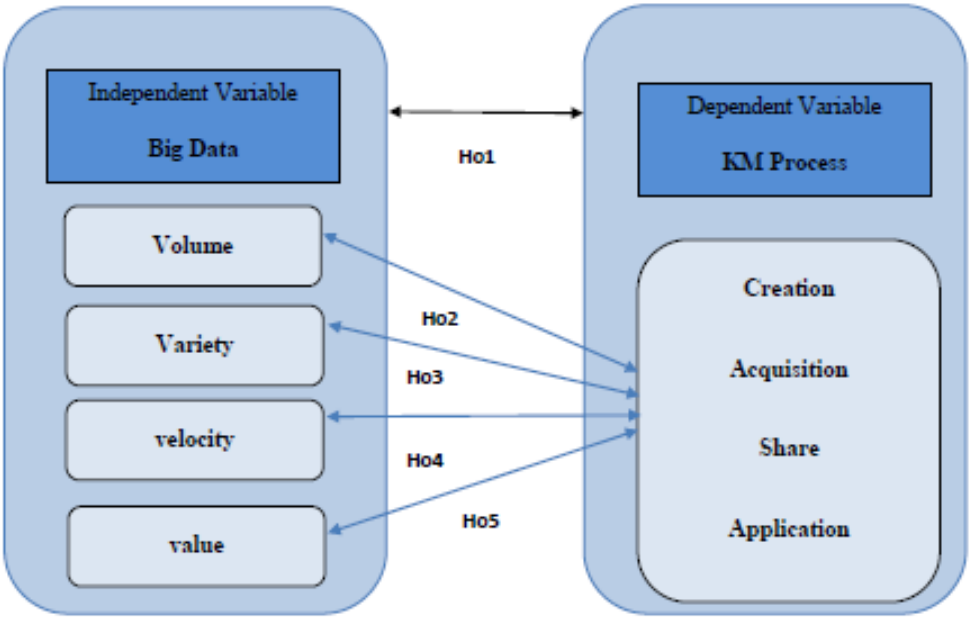


Figure 2.1: Relationship between big data and knowledge management

Source: ***(Alhinn & Rababah, 2018)***

The authors claim that the process of knowledge management includes knowledge acquisition, creation, transfer, sharing, and utilisation (Alhinn & Rababah, 2018: 10). Each element within the knowledge management process is motivated by data innovation, thus linking the knowledge management process to big data. In the above figure, the big data Vs are also illustrated as key components of the link between knowledge management processes and big data. Big data dimensions can, therefore, only be utilised and discovered if they are addressed within the knowledge management process. Alhinn and Rababah (2018: 23) recognise that big data identify hidden knowledge within the large data sets available, thus also assisting in the determining of strategies with the use of knowledge management processes (Alhinn & Rababah, 2018: 10). This also links to the fact that big data awareness needs to be orientated towards the process of knowledge management. Linking big data and knowledge management improves the capabilities of the environment in which the big data exists (Alhinn & Rababah, 2018: 23).

As knowledge management is recognised as a key driver of organisational performance (Ling *et al.,* 2008), it is an imperative for organisations to take full advantage of the value of knowledge which exists in the big data environment (Alhinn & Rababah, 2018: 25).

Alhinn and Rababah (2018: 25) advocate that emphasis should be put on individuals, procedures, and innovation within the organisation, as these three extremely important aspects assist with the leveraging of knowledge and thus the leveraging of big data. It was also recognised that even though the advantages of knowledge management and big data are known, they are not yet fully utilised in these environments, hence, the emphasis needed on aspects which can assist the motivation for big data and knowledge management processes to work hand-in-hand to be an advantage for the environment in which it exists (Alhinn & Rababah, 2018: 25).

## 2.3 Big data governance

**Data governance** (DG) is the overall **management** of the availability, usability, integrity and security of the **data** used in an enterprise. A sound **data governance** programme includes a governing body or council, a set of procedures and a plan to execute those procedures. **Data governance** is required to ensure that an organisation's information assets are managed formally, properly, proactively and efficiently throughout the enterprise to secure trust and accountability. Adopting and implementing **data governance** can result in improved productivity and efficiency. **Data governance** (DG) is usually manifested as an executive-level **data governance** board, committee, or other structure that makes and enforces policies and procedures for the business use and technical management of **data** across the entire organization. According to Seiner (2013: 15-16):

> A **data governance framework** refers to the process of building a model for managing enterprise **data**. The **framework** or system sets the guidelines and rules of engagement for business and management activities, especially those that deal with or result in the creation and manipulation of **data**.

A **data governance policy** is a documented set of guidelines for ensuring the proper management of an organisation's digital information. Such guidelines can involve **policies** for business process management (BPM) and enterprise risk planning (ERP), security, **data** quality and privacy. Seiner (2013: 15-16) explains that "**Data governance** is the formal orchestration of people, processes, and technology" that enables an organisation to leverage **data** as an enterprise asset. Raw **data** are largely without value, but it can become an organisation's most important asset when it is refined and understood.

The data steward in data governance is therefore responsible for keeping track of the data and knowing where the data came from, how to find it, and if it is trustworthy or not (Washington, 2018). The data steward would, therefore, be responsible for establishing a framework for data governance, which can assist in the establishment of data understanding, and to set data quality benchmarks (Washington, 2018). Furthermore, data governance will include the steward assigning accountability and ownership among data stakeholders, and the development of appropriate metadata (Washington, 2018).

Data governance can be defined as the functional co-ordination and definition of the processes, policies, standards, technologies and people within the environment to manage the data as an asset (Informatica, 2013: 3; Zhang & Yuan, 2016: 2). The governance of big data enables the availability and controlled growth of accurate, consistent, secure and timely data for better decision- making, reduced risk and improved organisational processes (Informatica, 2013: 3; Zhang & Yuan, 2016: 2).

Metadata is a crucial part of big data governance, and the metadata types applicable to big data governance in any environment are technical metadata, business metadata, and operational metadata (Informatica, 2012: 3). Each of these is described in more detail below.

Technical metadata include, for example, the technical information regarding the data, this being the source, column names, data type, and the data string. This enables the data to be identified by means of their technical aspects (Informatica, 2013: 3; supported by Zhang & Yuan, 2016: 2-7).

Business metadata includes the business context surrounding the metadata, the definition, the data stewards, and associated reference data (Informatica, 2013: 3; supported by Zhang & Yuan, 2016: 2-7).

Operational metadata include things influencing the use of the data, such as the date the data were last updated, the last time they were accessed and the number of times they have been accessed (Informatica, 2013: 3; supported by Zhang & Yuan, 2016: 2-7). With all the above in mind for the metadata governance of big data, the aspects regarding effectively governed metadata can be identified.

The first thing provided by appropriately governed data is visibility into how the data flow through the given environment and how the relevant big data life cycle is completed (Karel, 2013: 4).

The second aspect is the impact analysis and root cause analysis. Impact analysis includes the opportunity for users to see how a certain change may impact the environment before it happens. In turn, root cause analysis assists the organisation in identifying the root cause of a fault or problem within the organisation by means of problem-solving tools and techniques (Karel, 2013: 4).

The third aspect is that of governed metadata providing a common business vocabulary for the standardisation of technology within the environment, thus identifying an efficient organisational directory which will create a clear flow of communication and use of data terminology between units within the organisation (Karel, 2013: 4).

The fourth aspect of governed metadata is that it provides accountability regarding who is responsible for which terms and definitions of the given data within the organisation, thus identifying a clear line of responsibility regarding who has created what terms for data and who has made any changes or worked with any given data (Karel, 2013: 4).

The last aspect to consider in metadata governance in the organisation, is that it provides audit trails for compliance, meaning that any individual's work with any given data in the organisation is tracked and traced in order to create the authenticity and responsibility of who has worked with the data, and the transparency of who is responsible for which actions in the organisation. This also influences the flow and life cycle of the data within the organisation, and how they move and are made use of by each given individual (Karel, 2013: 4).

Furthermore, Karel (2013: 13), identified practical considerations for metadata management in an organisation. These practices are presented in ten steps, as follows:

1. Show up, start small and execute. The metadata steward should start on smaller projects of data governance and build from there, this ensuring project success.
2. Quantify everything. The steward should be able to justify their work regarding the metadata governance at any given time to illustrate transparency.

3. Set a focussed and reasonable scope. The steward should work in a timely and realistic manner on their projects and goals and not overwhelm themselves with too much at once as this may produce quantity in the end, but it hampers quality.
4. Get executive sponsorship. Sponsorship should not be obtained after the first failure of the attempted project, but before the project of data governance even begins. Success is achieved by ensuring that the appropriate resources are in place for efficient use.
5. Establish a data governance initiative. Such an initiative will ensure that the organisation will make use of metadata as a business benefit and not just an IT productivity tool.
6. Pack a high-value, low-complexity target as your pilot. When identifying a problem to solve, select one with a high return of value once it has been solved, thereby ensuring a value-added process in the activity log.
7. Assign owners and get business users involved. Data without a business context has no value. Any data within a business needs context, otherwise it is deemed to be useless.
8. Use both a carrot and a stick. A carrot is an incentive relating to data management. A stick can be considered as ensuring effective support from management regarding the data activity planned by data managers.
9. Tie metadata management to a business initiative. To ensure successful data governance, the business should consider attaching metadata management and data stewardship best practices to any upcoming projects as the projects are initiated through IT.
10. Identify a potential data crisis and be prepared for it when it occurs. Data stewards need to be prepared with a proposal in the case of the occurrence of a data crisis in the business. This does not necessarily mean that the data crisis will happen, but it means that the business is prepared for any unsuspected events in the future.

The practical considerations mentioned above may not guarantee the ultimate success of the organisation's (meta)data or data governance but will ensure long-term continuity and increase the chances of success of the data governance.

### 2.4 Risk and big data management

The risks of big data include privacy breaches and data security, inconsistent access and continuity, resistance of big data providers and populace, fragmentation of approaches across

jurisdictions, resource constraints and cutbacks, and privatisation and competition (Kitchin, 2015: 4).

Risk management can be defined as the forecasting and analysis of risks together with the identification of procedures that can be undertaken to minimise or avoid these risks. Veldhoen and De Prins & Veldhoen (2014: 5) explain how risk management faces new challenges in response to the need for more detailed data. Big data address the need, but the question is: How will organisations manage the big data technologies that can address the risk challenges better? (Veldhoen & De Prins, 2014: 5). In this regard, the responsibility of the data steward is to forecast and analyse risks associated with big data and their management.

Furthermore, Veldhoen and De Prins (2014: 5) state that comprehensive and diverse real-time data may be the solution to the challenges, by improving the monitoring of risk, risk coverage, and the predictive power of risk models. Veldhoen and De Prins (2014: 5) discuss how big data technologies will also allow the development of models that may support risk management decision-making for the better. These big data technologies can also accommodate new challenges and demands for any given situation relating to risk (Veldhoen & De Prins, 2014: 5). Lastly, big data help risk management through the organisation having better predictive power, anti-money laundering with real-time actionable insights, organisational risk being addressed faster and more accurately, and the fact that big data technologies may assist in covering all aspects that are related to risk management, thereby, helping to mitigate current and potential risks (Veldhoen & De Prins, 2014: 6-9).

## 2.5 Intellectual property rights

When considering how big data, or a data set may be protected by intellectual property, Solove *et al.* (2014) explain how data are protected by patents, copyright, and trade secret licences. Big data may be protected under the same types of licences but the problem is that data cannot be re-used on a large scale if there is insufficient information describing its provenance (Solove *et al.*, 2014). The intellectual property system will have to be modified, otherwise big data and their capabilities will never be realised as the intellectual property system at present is aimed at incentivised technological disclosures, while big data are aimed at society (Solove *et al.,* 2014).

Tyhurst (2018: 1) discusses how big data are becoming more prevalent within the intellectual property sphere and this should be noted more as they are used to allocate research funding, the determining of patents, and the protection or litigation of works. As big data develop in the intellectual property environment, as is the case with its influence on that environment.

## 2.6 Big data and development

According to Ismail (2016), "The bigger and better data gets in the developing world, the easier it will be to improve lives." Successful and appropriate big data stewardship can only enhance and improve business operations, services within the country, and reliable insight into situations which could previously not be resolved.

With regard to big data in developing countries, Hilbert (2016: 142) gives advice on how to apply big data for development. The first component of the application is to track words. When large datasets of words are analysed accurately and efficiently, activities and actions can be predicted within the given environment (Hilbert, 2016: 142). The second aspect of big data application for development, is tracking locations, meaning that geographic big data, if analysed properly, can be used to track travelling sequences, locations and migration patterns (Hilbert, 2016: 143; Manyika *et al*., 2011). The third aspect to consider is the tracking nature, meaning that if the big data are analysed and used appropriately, they may assist in reducing the uncertainty created and this may thus optimise the performance, mitigate risks, and improve emergency responses (Hilbert, 2016: 145). The fourth application aspect is tracking transactions, which can also be done appropriately. These transactions are footprints of social interactions and can be used as a cheaper way of measuring poverty levels in a country in real-time at a fine-grained geographic level (Hilbert, 2016: 146; Helbing & Balietti, 2010). The fifth application aspect is tracking behaviour, meaning that big data can be used to track abnormal behaviours that may also lead to different types of variations of behaviour tracking. These influencing types are environmental conditions, medical errors, overuse and oversupply, and biased judgements (Hilbert, 2016: 146). The sixth application of big data for development is tracking production. This includes reporting and identifying economic production, which can reveal competitive advantages in the environment, and perhaps illustrate opportunities for development (Hilbert, 2016: 147). The last big data application for development is tracking other data. This includes tracking the big data of financial, economic

and natural resources, educational aspects, waste, and expenditure and investment. Analysed big data are the sources of all these aspects (Hilbert, 2016: 148; Manyika *et al.,* 2011).

Ohri (2015) has also identified six opportunities for big data for development. These opportunities are improved financial services, increased agricultural opportunities, improved education standards, improved healthcare, reduced corruption, and monitoring carbon consumption (Ohri, 2015).

Even though the above-mentioned literature identifies big data for development, it does not necessarily identify big data stewardship. It is further noted that 'big data,' as a concept, may already be recognised in developing countries, but "big data stewardship" is an even newer term. Developing countries need to recognise that successful application of big data can enhance development in their countries.

### 2.6.1 Big data stewardship in Africa

As 'big data' is a relatively new concept, 'big data stewardship,' as noted above, is also a new concept in South Africa and the rest of Africa. If developing countries can utilise big data and big data stewardship opportunities, they can only reap benefits from this implementation.

Big data are, however, still a challenge for developing countries – more specifically those in Africa. The use of big data enable developing countries to determine the extent of problems such as food, water, energy, electricity, sanitation, and education. Once the constraints have been identified, reliable solutions to the problems can be found (Mutuku, 2016). Mutuku (2016) further states that studies have identified a great deal of inequality around the world in researchers' access to infrastructure, software, skills, and networks, including the support necessary to interpret and share big data. These inequalities should be addressed to avoid another divide occurring between high-income and low-income countries. Therefore, Mutuku (2016) suggests that African countries should consider big data as an opportunity with societal benefits. This opportunity can only be grasped if certain conditions are met. The first condition pertains to certain data factors, which include the fact that the large amounts of data available must be available for a large number of users and not only for a few. The next factor is the human factor, which includes the ability of humans to analyse the given data. Lastly, the political factor includes the willingness to make decisions based on accurate data (Mutuku, 2016; Kshetri, 2016: 23).

Low production costs have made smart devices more affordable for even the poorest communities in Africa. Data scientists have discovered a strong relationship between smart device usage/data usage and food consumption in developing countries, which is proved by the fact that the more data citizens use on their phones, the more likely they are to purchase higher quality foods. However, where less data is used, the more likely it is that the food, that is purchased, is unhealthy (Ismail, 2016). Having said that, Ismail (2016) also reports on another initiative for big data in Africa where it became evident that, in developing countries, big data are being used to positive effect, such as saving lives because researchers can now use smart devices to gather, analyse and send data. Accordingly, this increase in smart device usage has led to an increase in data usage, which, in the end, can create an opportunity for data scientists and stewards to analyse the data more precisely and accurately It was also disclosed that hunger can be prevented by being prepared, with which big data can assist (Ismail, 2016).

From the above examples, we can see that big data are a key prerequisite of an enhanced and enriched life in African countries, therefore, it is a matter of the big data being used and analysed correctly by data scientists. Although there have been successful initiatives with regard to big data development and application in Africa, big data stewardship is still a fairly new occupation and concept for these countries. Accordingly, big data have become an integral part of any research and funding application for big problems in Africa. One reason why there is a significant difference in development in big data applications between developed and developing countries is that developed countries make use of big data stewards, but African and developing countries are only at the beginning stages of harnessing the role of big data (Peters, 2017). In other words, Africa may be applying big data for development, but big data stewardship is also needed for this development not only to take place, but also to be sustainable and enduring.

### 2.6.2 Big data stewardship in South Africa

South Africa is still a developing country, even though it is often regarded as the most developed country in Africa. As with big data stewardship elsewhere in Africa, South Africa has also generated opportunities and initiatives to promote data stewardship, but application and adoption will make the difference. An initiative that can be discussed is that of the Data

Intensive Research Initiative South Africa (DIRISA), which has recognised that big data are an integral part of research and development in any country and if South Africa has a desire to transform into an information and knowledge society, the country will first have to equip people with the necessary skills to meet the requirements that new technology presents (Van de Groenendaal, 2016). Furthermore, Van de Groenendaal (2016) points out that skills needed to work with and to understand big data, this being the reason for the launch of DIRISA that has formulated three objectives that they hope to achieve. The first is the research and development group within the institute to identify new research, such as distributed and streaming machine intelligence and convex optimisation. The second is a training programme for individuals needing to equip themselves regarding all aspects of big data. Thirdly, DIRISA wants to advocate data-intensive research to promote sound data stewardship practices, to develop the expertise of the given individuals, and to co-ordinate data-intensive research activities (Van de Groenendaal, 2016). DIRISA further states that when doing research regarding the development of the initiative, business and industry were consulted and it was noted that individuals, qualified in handling big data, are scarce and in great demand (Van de Groenendaal, 2016). This is one of the larger drivers for the DIRISA initiative.

With the detail provided above in mind, one notes that South Africa still has much to do to develop the skills to implement big data stewardship. There is not only a shortage of qualified individuals, but also a shortage of training opportunities, which creates a downward spiral. Therefore, more opportunities need to be created for individuals to be trained to become data stewards, to meet the urgent demand for these qualified individuals.

## 2.7 Big data life cycle models

Gaining a good understanding of the data life cycle allows the curator to understand the nature of the management responsibility (Jarosciak, 2017). The life cycle can be described as a picture worth 1000 words, and is, therefore, probably the most important component of making sense of the big data stewardship challenge. As explained by Jarosciak (2017), big data faces constant challenges like distributed processing, semantic integration, association mapping, and timeliness. As the concept of ‘big data’ is constantly evolving, it is difficult to pinpoint a definitive big data life cycle model. For this reason, one can compare different big data life cycles and note the similarities, acknowledging that a new definitive model may be

published at any time. Data governance (Chisholm, 2015) is a perfect example of how the big data life cycle can be thought of, in terms of, "What would happen if we could ride on a piece of data as it moved through the enterprise? What new experiences would the piece of data have? What phases would it pass through?" This is a perfect way to think of the big data life cycle and what could and does happen to data in the life cycle.

### 2.7.1 Phases of a typical big data life cycle model

A big data life cycle typically entails a description and a visualisation of the different phases through which the data move from the starting point to a point where the data are no longer of immediate use. According to Pouchard (2016:180), big data life cycle models present structures and frameworks for organising all tasks and activities related to the management of big data within any environment. The big data life cycle means the data are subjected to a process of input, a process and an end product is delivered as the output. The cycle is a way in which tasks can be communicated to the designated audience, which include all the participants, namely researchers, data managers, curators, repository specialists, librarians and project managers (Pouchard, 2016: 180). There are several life cycle models and it is difficult to choose the one above the other.

A life cycle can be applied in any context, as it demonstrates each stage that must be managed efficiently for a successful outcome. It is important to note that each phase of the big data life cycle has its own characteristics.

Big data do not arise spontaneously, it must be created (Jagadish *et al.,* 2014: 89). The first phase of the big data life cycle is creation. Data creation refers to the actual starting point of big data collection. It includes the initial planning as well as the collection of data that are created and sourced to work through the life cycle (Jagadish *et al,.* 2014: 89). Big data do not just arrive in the cycle and come into existence all of a sudden, it is a record of the underlying interest in the data that have thus generated the big data which now exist and to which value can be added as the data move through the cycle to become more relevant and useful.

The second phase to note is that of capturing the big data, otherwise known as data acquisition. Pouchard (2015: 185) discusses how the acquisition activity reflects how data are produced, generated, and ingested in the data capturing process. In other words, data acquisition is the ingestion of data which have already been created and already exist that are

produced by an organisation because of an underlying interest in and need for the data (Chisholm, 2015). Data capture or data acquisition entails the capturing of millions of data points which become big data with the intention of sending the data through the big data life cycle for further evaluation and cleaning so that the data become more useful and more valuable.

Data curation is the third phase of the big data life cycle. Data curation refers to the management of the data to enable more complete and high-quality data-driven models for the organisation at hand (Freitas & Curry, 2016: 89). The data curation processes consist of activities such as content creation, selection, classification, transformation, validation, and preservation (Freitas and Curry, 2016: 88). Furthermore, data curation entails providing methodological and technological data management support to address data quality issues so as to achieve the maximisation of data usability (Freitas & Curry, 2016: 87).

The next phase of the big data life cycle is the transfer stage. The transfer stage includes how the big data are prepared for analysis. It involves reformatting, cleaning, and integrating data sets into a desired format for further stages in the life cycle (Pouchard, 2015: 186).Furthermore, it involves data wrangling within the life cycle where iterative data exploration and transformation takes place, thereby ensuring that the data gains value (Pouchard, 2015: 187).

The next phase of the big data life cycle is data storage that includes the copying or moving of data to an environment where the data can be stored efficiently and accessed easily when needed again. More simply, data storage entails the place where the data are stored and preserved easily and accessed easily when needed in future. This place also precludes the maintenance needed for the data, and the opportunity for data to be restored when necessary (Pouchard, 2015: 186).

Data analysis is the next phase. This includes the domain in which researchers or big data managers work to ensure that the data is of the correct quality and is prepared correctly to meet the needs of the environment in which it exists, thereby ensuring that the big data are valuable and are in the correct format to be processed and used efficiently (Pouchard, 2015: 186). With the correct analysis techniques, big data analysts can ensure good results without being overwhelmed by the volume of data present (Jagadish *et al.* 2014: 90).

The last phase of the big data life cycle is big data visualisation. Big data visualisation can be described as the output of the data life cycle process, as it involves the manipulation of the data into a format, table, pattern or correlation, which is understandable for the users of the now valuable big data (Jagadish *et al.,* 2014: 90). Furthermore, big data visualisation entails the packaging of the end product of the life cycle so that the data are understood and used easily. The visualisation of the data may, however, differ in each situation, as big data are discipline-specific and may need to be produced in different packaging to be understood by the target market (Jagadish *et al.,* 2014: 90).

### 2.7.2 Big data cycle models used in business vs Big data cycle models used for research

When evaluating big data life cycles for different environments, it is important to remember in which environment the life cycle exists, namely, a business environment or a research environment.

In contrast to a researcher and curator's big data life cycle, the business environment needs to make use of a big data life cycle for business analysis, which is applicable to everyday business procedures, tasks and activities which need fulfilling by means of big data. A big data life cycle which can be applied to a business environment, is presented in Figure 2.2 below.

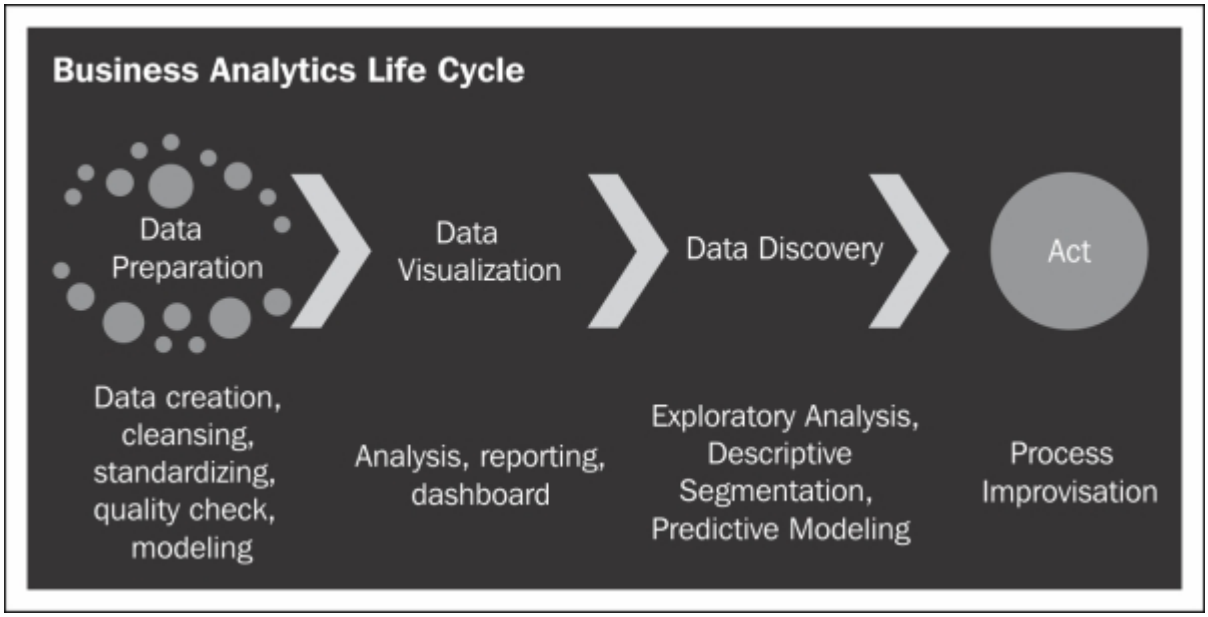


Figure 2.2: Business analytics life cycle (Patil and Thia, 2013)

In the business environment, and as seen in Figure 2.2, the first phase links to data preparation which includes activities such as data creation, cleansing, standardisation, quality control, formatting and modelling. The second phase is data visualisation and contains the activities data analysis and data reporting (Patil & Thia, 2013). The next stage of the business life cycle is that of data discovery, which includes activities such as exploratory analysis, segmentation, and predictive modelling. Lastly, the data are acted upon in accordance with the needs and desired outputs (Patil & Thia, 2013). Big data being discipline-specific ,obviously influence the activities in the business life cycle, as these activities change in accordance with the environment in which they exist, meaning that one activity may be more important to carry out within the life cycle than another, which may be different for another business environment, this all depending on the businesses' desired outcomes and need for data.

Regarding a big data life cycle which meets the needs and outcomes of a researcher, the model presented below is perhaps more appropriate.

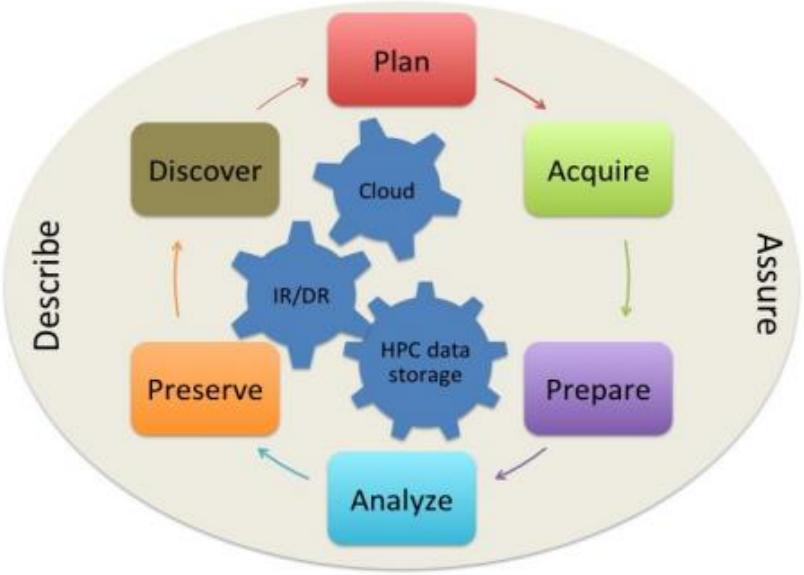


Figure 2.3: Big data life cycle model (Pouchard, 2015: 184)

Figure 2.3 illustrates the big data life cycle, which is more appropriate for research-intensive work, meaning that the processed and outcomes with regard to the data are different from the business big data life cycle. The two life cycles have different end goals and needs concerning the data. The research-intensive big data life cycle includes all the characteristics

that have been discussed, such as planning, acquiring, preparing, analysing, preserving, and discovering (refer to section 2.7.1 for a description of each of these). The research-intensive big data life cycle also includes the assurance and description activities. With this in mind, the central cogs in Figure 2.3 represent the storage infrastructure, which could allow the data to be stored in the cloud, or in an institutional repository, a disciplinary repository, or a high-performance computing centre storage facility (Pouchard, 2015: 184).

The intent behind the two big data life cycles illustrated, is the same: use of a standard pattern to move data from the inception of the process to the end. It is important to consider which big data life cycle is to be used, by evaluating the environment in which the life cycle needs to operate, and the data needs which are to be fulfilled.

### 2.7.3 Big data life cycle model

From the life cycle commonalities recognised above, a generic life cycle can be formed, which will highlight the most important commonalities for big data. The commonalities between the life cycles discussed above were data creation, data acquisition, quality control, data analysis, data visualisation, data preservation and storage, and data re-use. In the given particular order, the life cycle can then be illustrated as follows in Figure 2. 4.

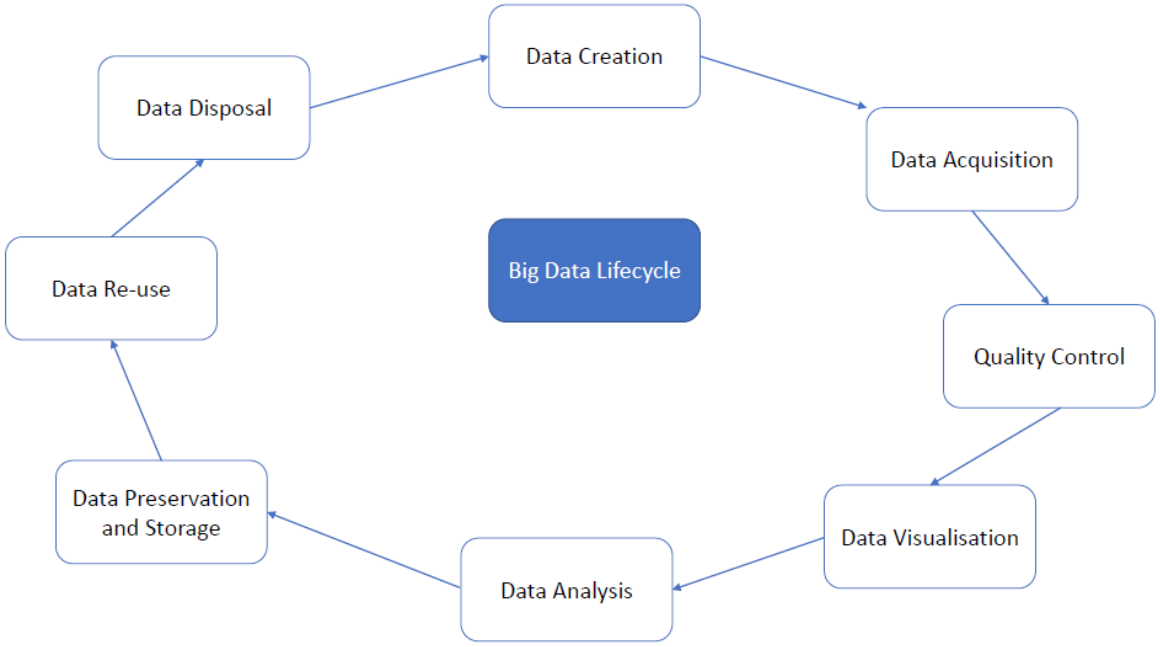


Figure 2.4: A working model of the typical big data life cycle (Source: author's own illustration)

The components shown in Figure 4 have already been discussed in section 2.7.1. It is also important to understand the order of the life cycle and the purpose of that order. Data creation is the first stage of the life cycle. This is where the data are generated for further use. The second phase of the life cycle is data acquisition. This is when the necessary data are collected purposefully from the data which were created. This is a filtering of the data, which leads to the third stage of the life cycle. The third phase is quality control, when the purposively collected data go through quality controls to ensure that the data meet the necessary standards before further use. The fourth phase is data analysis. This gives a sense of understanding of the data, its purpose and future use. The fifth stage is data visualisation, which creates a better understanding of the now purposeful data. Data preservation and storage is the sixth phase, when the useful data are stored and preserved successfully for future use. The algorithms and visualisations used to understand the data must also be preserved. The last phase is data re-use, which happens when the data can serve a new purpose, when it may go through the life cycle again, before it is disposed of or archived.

In the next two sections, two existing curation models are examined to identify and superimpose the curation activities onto the big data working model. The first of the curation models investigated is the well-known DCC curation model.

### 2.7.4 The DCC curation model and big data

The DCC (Digital Curation Centre) curation model is illustrated below in Figure 2.5.

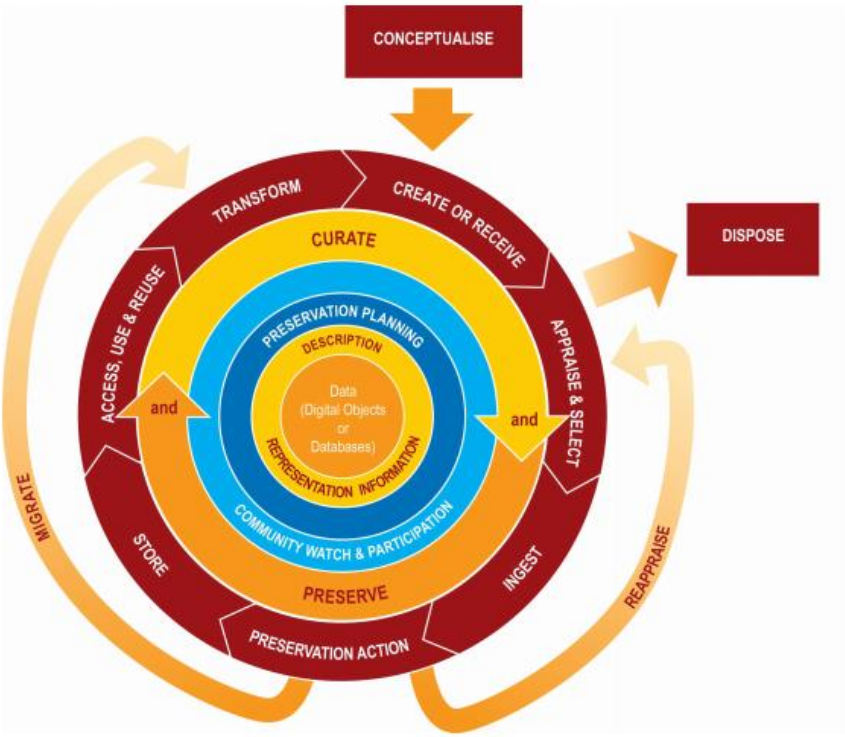


Figure 2.5: The DCC curation life cycle model (Higgins, 2008: 136)

The DCC curation life cycle model provides a graphical illustration of the stages required for the successful curation and preservation of data (big or long tail) through the iterative curation life cycle. It also defines the roles and responsibilities needed to build a framework of standards and technologies to be implemented (Higgins, 2008: 136). The question is whether these activities would still be appropriate when applied to big data. Would a data steward still have a role to play when the dataset is really big? To answer this question, it is necessary to list the activities associated with each of the outer circle phases provided in the model. Table 2.1 on the next page, summarises the detail.

*Table 2.1: DCC life cycle model checklist for data stewardship*

| Life cycle stage | Stewardship activity | Applicability to big data stewardship |
|---|---|---|
| Data conceptualisation | • Equate data curation with good research<br>• Know what the funding body expects of the data, and for how long<br>• Determine intellectual property rights, and ensure that they are documented<br>• Identify expected publication requirements<br>• Identify and document specific roles and responsibilities as early as possible | Applicable |
| Data creation | • Know for whom the data are created and for what it will be used<br>• Identify data protection requirements in the course of the research<br>• Agree on the standards to be used, and communicate accordingly<br>• Identify data quality metrics as soon as possible, and ensure that they are communicated<br>• Work together with researchers and information managers<br>• Be realistic about what is sufficient and what is ideal | Applicable |
| Data selection and appraisal | • Start with selection and appraisal as early as possible<br>• Plan for what will be needed to keep supporting research findings | Applicable |

| Life cycle stage | Stewardship activity | Applicability to big data stewardship |
|---|---|---|
| | • Know for whom the data are kept, and what they will be doing with them<br>• Know what should be disposed of to meet legal requirements<br>• Ensure that the data meets the minimum quality assurance metrics<br>• Re-appraisal can take place before ingestion, to review what has been done and what needs to be deposited for long-term use<br>• Work with researchers and information managers to formulate policies<br>• Appraise for the current status and the future status | |
| Data ingestion and storage | • Make use of archival standards for hierarchal data description<br>• Know about repository policies<br>• Ingestion may not mean data deposit, but could mean moving the data to a curated environment<br>• Make the ingestion process as straight-forward as possible<br>• Decide who is responsible for the final aspects of the data quality assurance at the point of deposit<br>• The level of data quality and cleaning must be assessed by fitness for purpose | Applicable |

| Life cycle stage | Stewardship activity | Applicability to big data stewardship |
|---|---|---|
| | • Get a formal receipt or an informal acknowledgment for closure and transfer of stewardship | |
| Data preservation | • Know what users should do with the data<br>• Make sure that the users of the data carry out preservation actions<br>• Be critical when reviewing the best practices for the data and the recommended approaches<br>• Document preservation actions, so that users know what has been done to the data over time | Applicable |
| Data access and re-use | • Know what users should be able to do with the data, and for how long<br>• Communicate significant properties of data<br>• Ensure that any restrictions on and access and use are communicated and respected<br>• Ensure that enough context is provided for the data to be located and used | |

The DCC life cycle starts where the research data life cycle ends – when the researcher hands the data over for curation. The model has detailed stages through which the curator or data steward must work to process the data into a product for the secondary data user. The curator can use the life cycle as a guideline to ensure that the data are curated effectively. The DCC life cycle stages include the digital objects and the databases, the assigning of metadata to the data, preservation planning, community watch and participation for development, curation and preservation promotion, conceptualisation, creating of metadata and receiving of data, evaluation by means of appraisal, ingestion of the data, long-term preservation of the data, data storage, data accessibility, data transformation, data disposal, reappraisal of data, and lastly, data migration.

### 2.7.5 The UK data archive model and big data

The life cycle shown in Figure 2.6 is the research data life cycle, which has been annotated from the data archive cycle provided by the UK Data Archive. In terms of this model, “it is important to note what activities the curator is responsible for at each of the various stages of the life cycle.” The life cycle illustrates how curation also involves the re-use of data, the processing of data, the analysis of data, the preservation of data, and lastly the access responsibilities associated with the data (Van Den Eynden, 2012). The processing of data that curation includes are the entering of data, the digesting and translating of data, the checking and cleaning of data, the managing and describing of data, and the anonymising of data when necessary (Van Den Eynden, 2012). The analysis of the data pertaining to the curation, includes the interpreting of the data, the deriving of the data, the producing of research outputs, author publications, and the preparation of the data for preservation (Van Den Eynden, 2012). The preservation of data, for which curation is primarily responsible, includes the migrating of data, the backup and storage of data, the creation and documentation of metadata, and the archiving of data (Van Den Eynden, 2012). The right of access that involves curation activities includes the distribution of data, the sharing of data appropriately with the right consumers, the controlling of access to the data, the establishment of copyright of the data, and the promotion of the data (Van Den Eynden, 2012). Lastly, the reusing of the data which involves curation activities includes following-up on research, the establishment of new research, the undertaking of research reviews, the analysing of findings, and the teaching and learning associated with the use of the data (Van Den Eynden, 2012).

When considering the above involvement of curation activities within the life cycle and big data, one can note that the activities will not necessarily change but may become more intricate or may be on a larger scale. The primary activities of curation in handling the data remain the same.

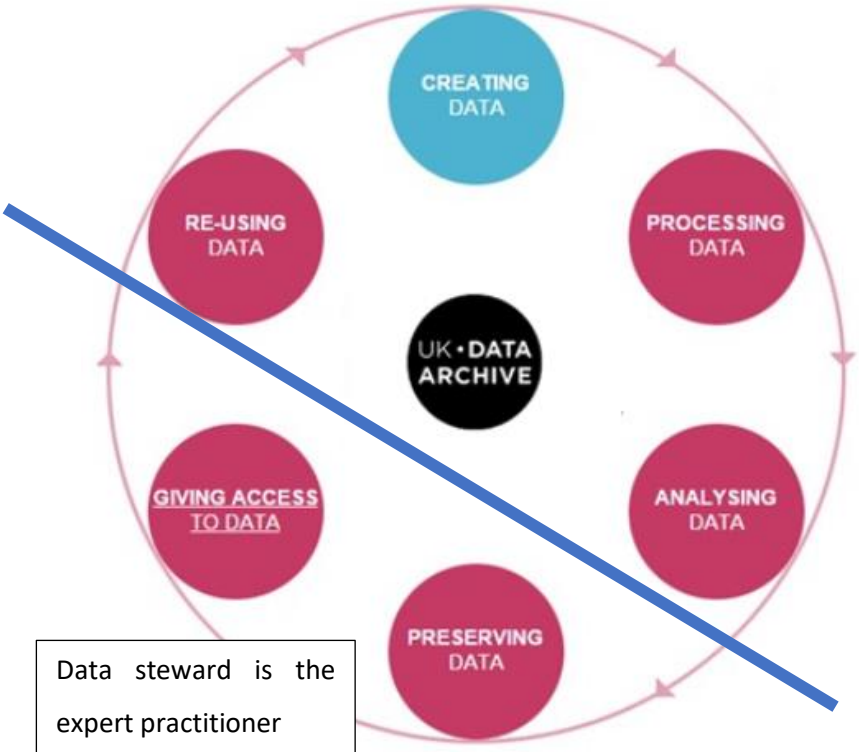


Figure 2.6: The research data life cycle model of the UK Data Archive (Van Den Eynden, 2012)

*Table 2.2: UK Data Archive table and big data stewardship*

| **Life cycle stage** | **Stewardship activity** | **Applicability to big data stewardship** |
|---|---|---|
| Creating data | • Design research<br>• Plan data management<br>• Plan consent for sharing<br>• Locate existing data<br>• Collect data<br>• Capture and create metadata | Applicable |
| Processing data | • Entering of data<br>• Digesting and translating of data<br>• Checking and cleaning of data<br>• Managing and describing of data<br>• Anonymising of data when necessary | Applicable |
| Analysing data | • Interpreting of the data<br>• Deriving of the data<br>• Producing of research outputs<br>• Author publications<br>• Preparation of the data for preservation | Applicable |

| Life cycle stage | Stewardship activity | Applicability to big data stewardship |
|---|---|---|
| Preserving data | • Migrating of data<br>• Backup and storage of the data<br>• Creation and documentation of metadata<br>• Archiving of data | Applicable |
| Giving access to data | • Distribution of data<br>• Sharing of data appropriately with the right consumers<br>• Controlling of access to the data<br>• Establishment of copyright of the data<br>• Promotion of the data | Applicable |
| Re-using data | • Following-up on research<br>• Establishment of new research<br>• Undertaking of research reviews<br>• Analysing of the findings<br>• Teaching and learning associated with the use of the data | Applicable |

### 2.7.6 A data stewardship intervention model for big data

From the information provided above, it was possible to compare the models and consolidate the activities that could be associated with big data stewardship. The table following compares the models and lists the identified stewardship interventions, actions or responsibilities.

*Table 2.3: Life cycle model comparison, with stewardship activities added*

| **Working model of big data life cycle** | **DCC's curation life cycle** | **UK Data Archive life cycle** | **Associated stewardship interventions** |
|---|---|---|---|
| | Conceptualise | | Understand data, as a whole, with all relevant components before entering into life cycle |
| Data creation | Creating data | Data creation | Plan data management and all related components. Know what data are needed and for what they are needed |
| Data acquisition | | | Obtain and capture relevant data to suit relevant needs |
| Quality control | | | Ensure that the data collected are filtered through a quality control process so that only the appropriate data are forwarded through to the life cycle |
| | Appraise and select | | Appraise the relevant data and go through a selection process to satisfy the needs |
| | Ingest | | Access and import data for immediate use or storage in a database |
| | Processing data | Processing data | Process the relevant data to create a more finished product |

| **Working model of big data life cycle** | **DCC's curation life cycle** | **UK Data Archive life cycle** | **Associated stewardship interventions** |
|---|---|---|---|
| Data visualisation | | | Assist users in understanding the data by illustrating the data in a physical platform |
| Data analysis | Analysis | | Analyse the data and describe the data in a way that users understand fully |
| Data preservation and storage | Preservation and storage action | Preserving data | Preserve the data for immediate use or future use, as necessary |
| | | Giving access | Give access to workers and users of the data for appropriate use |
| Data re-use | Access and re-use | Re-use data | Ensure that the data are preserved correctly, and in a form to be re-used when necessary |
| Data disposal | | | Identify data which can be disposed of responsibly if necessary. This includes irrelevant and unused data |

The comparison then led to an updated life cycle model, which includes the stewardship activities that would need to be unpacked when identifying the skills to develop. Stewardship activities identified from literature are illustrated in Figure

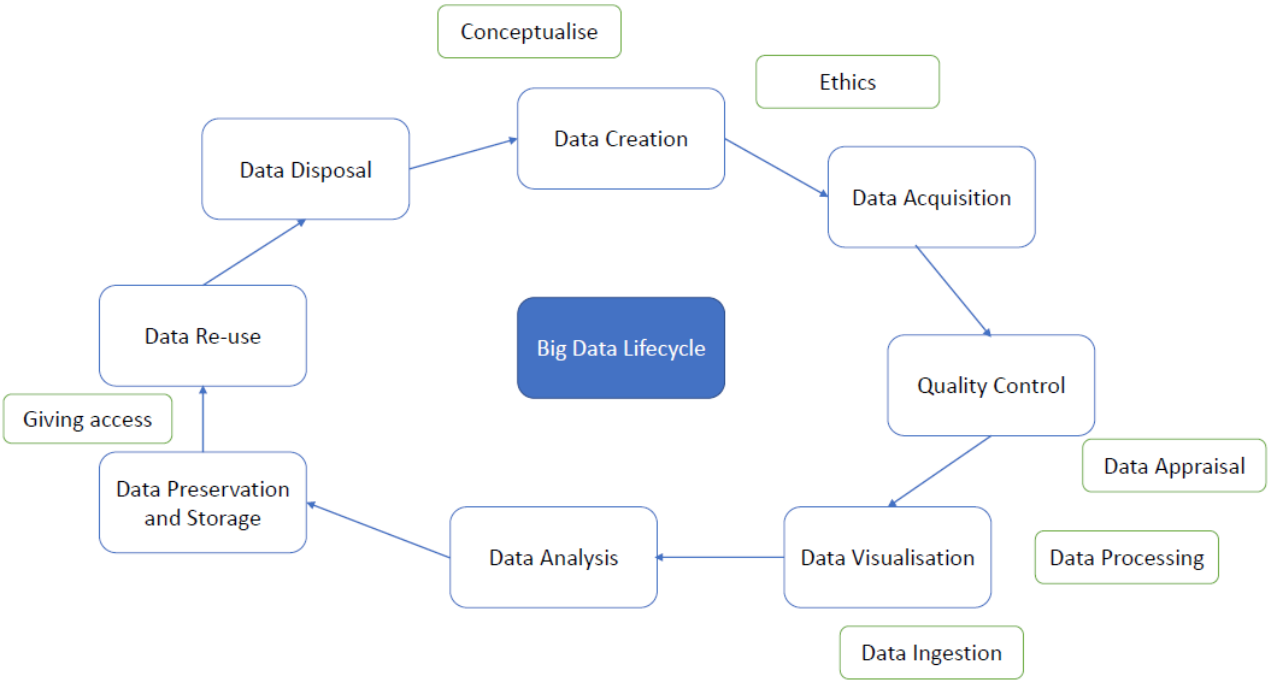


Figure 2.7: An intervention model for big data stewardship

Having identified the data management interventions and at what phase of the big data life cycle these should take place, it is necessary to take a closer look at the responsibilities of the data steward. Primarily, data stewards should ensure that the data management interventions or activities take place. Section 2.7 provides a more detailed explanation.

## 2.8 Summary

To summarise, this chapter acts as a foundation for the rest of the study. The literature, reviewed, assists in the development of the study and aids in the identification and understanding of the big data stewardship education requirements. Because 'big data' is a relatively new concept, it is expected that a variety of efforts have not yet converged to provide a *de facto* standard for training big data stewards. This research can, therefore, not only assist with the development of the concept of 'big data stewardship' itself, but should also encourage the development of big data stewardship in South Africa.

# Chapter 3

## 3. Literature review – Responsibilities, competencies and skills

### 3.1 Introduction

This chapter will act as the continuation of Chapter 2 that clarified and discussed the concept of 'big data.' while this chapter will discuss and present the roles and responsibilities identified for big data stewards, as well as the necessary skills needed to be a successful big data steward. This chapter will illustrate the content theoretically, and several matrixes and table formats were developed to aid with understanding this phenomenon. It is important to note that this chapter is closely linked to Chapter 2 as the content and literature reviewed of Chapter 2 have a direct link with this chapter.

### 3.2 Roles, responsibilities and challenges for big data stewards

The data steward should take ownership and control over the curation or data management interventions. The steward should have control over, oversee, and be responsible for, the curation process, which runs parallel to the data life cycle. However, before discussing the authority and ownership responsibilities of the data stewards, the specific roles of the data steward should be discussed.

#### 3.2.1 Roles

The first role of a data steward is to be the champion who facilitates and manages the appropriate and efficient use of the big data which fall within their domain (Brubaker *et al.* 2014: 4160). The data steward, therefore, plays the role of a collaborator and builds relationships with other data stewards to ensure that the data are associated with the correct metadata (Brubaker *et al.* 2014: 41-60). The data steward needs to be a trainer and is responsible for providing effective and accessible training, so that all the team members know how to manage the relevant data. The steward also plays the role of a quality assurer and is responsible for the development and promotion of data quality standards for data entry and the reporting of the specific data (Brubaker *et al*., 2014: 41-60). Lastly, the data steward is a gatekeeper because the steward is responsible for managing access to the data in a way which is consistent with the domain's regulations and the institutional data access philosophy. The

steward should also designate and delegate responsibilities to custodians who can take on some of the data management tasks (Brubaker *et al.* 2014: 4161). These responsibilities are linked to data governance (see section 2.3), risk management (see section 2.4), intellectual property rights (see section 2.5) or responsibilities such as training, managing access, quality control, applying FAIR principles and data ownership. Each of these is discussed in section 3.2.2 below.

### 3.2.2 Responsibilities

The identified responsibilities of a data steward will now be separated and discusse3d individually in order to clarify what each responsibility is, as well as how the responsibility may br fulfilled by the data steward.

#### 3.2.2.1 Training

A large part of data stewardship is to facilitate relevant training for stakeholders. The data steward needs to be trained first to be able to fulfil the tasks required of a steward. This includes being capable of training other stewards or other professionals who need to work appropriately with the data.

Training topics would be varied and broad – from things such as naming conventions, versioning, and labelling (Malinowski, 2016: 31-48; Segment, 2018) to the implementation of discipline standards, the use of national data infrastructure, and ethical data collection. A large part of training is naming conventions. When considering naming conventions and version control for the present data, it is important to name and control the data according to the environment in which it exists, or the environment in which the data should exist. Malinowski (2016: 30) states that naming conventions should make the tasks of a data steward easier. A file naming convention should thus be descriptive and consistent. Accordingly, aspects such as a unique identifier, a project or research data name, the conditions in which the data exists, the experiment itself, the date of the file properties, and the version of the data should be considered (Malinowski, 2016: 31). With regard to the data being consistent, it should maintain its order and it should always include the same information, to ensure the consistency of the data file naming conventions (Malinowski, 2016: 32). The data steward should assist with standardising the file naming conventions and in

training those involved to use the file naming conventions appropriately. This will ensure the successful implementation and use of the file naming convention.

Furthermore, the naming conventions of the data should include the data steward making sure that the instrument in use for naming conventions can be set up with the actual file naming convention needed, as this is in the end supposed to make less work for the data steward when working on file naming conventions (Malinowski, 2016: 35).

The data steward should also check for file naming conventions, which already exist within the given discipline to ensure the needed consistency. Segment (2018) notes three characteristics of naming conventions which should always be considered: the convention should be consistent, files should be easily found using naming conventions, and it needs to have clarity, meaning that if appropriate standards are set, all the team members work on the same understanding of the naming convention.

Version control helps the data steward to keep track of the data. Therefore, file versioning should take place on the data files and analysis files, script files, and programmes (Malinowski, 2016: 38). Project documentation and files also need to be versioned appropriately, by always saving new versions of the data and establishing consistent conventions for the new and older files (Malinowski, 2016: 41). Accordingly, to ensure consistent and clear file versioning, the data steward should use ordinal numbers for larger version changes, and decimal numbers for smaller changes. Making use of dates also assists with distinguishing between successive versions (Malinowski, 2016: 43).

It is also important for the data steward to avoid the imprecise labelling of new and older files, meaning that older versions of files should be put into a separate folder. The steward should consider if obsolete versions of the file should be kept or discarded (Malinowski, 2016: p45). Lastly, the data steward should create a version table alongside the data files to track history and activity and should make use of built-in capabilities and software for automated standardised naming conventions (Malinowski, 2016: 47). To sum up, the data steward should save new versions, establish a consistent convention, document the convention, and consider the version control needs of the data (Malinowski, 2016: 48).

With all of the above-mentioned regarding training, it is important for the data stewards to attend the training necessary firstly to maintain the skill set of a big data steward, but also to be able to train others who will be working with the data.

#### 3.2.2.2 Managing access

Managing access includes repository implementation, licensing and security, but it is also about proper citation, persistent identification and measuring impact. When considering access data, data citation is an important attribute. Rauber *et al*. (2015) state that the goals of data citation should be to create arbitrary views of data, from a single record to an entire data set, in a precise manner. The data citation should also allow the data steward to cite and retrieve data as they exist at a specific point in their life cycle, whether the current database is static or dynamic (Rauber *et al*. 2015). Lastly, data citation should be stable across all different types and versions of technology and technological changes, to create the consistency necessary for the data to be used efficiently (Rauber *et al.* 2015).

Certain steps can also be considered when working with data citations, the first being the preparation of the data and the query store, meaning that data versioning should take place along with ensuring that the data have been time-stamped, and storing any queries associated with the metadata so that these queries can be re-executed in future (Rauber *et al*., 2015). The second step is identifying specific data sets persistently, which includes considering query uniqueness, ensuring that sorting of data sets is unambiguous and reproducible, setting up a standardised result set verification system, timestamping queries (when they are resolved compared to when they were made), query identification, storing query metadata, and creating automated citation texts (Rauber *et al*., 2015). The third step is resolving query identifications and retrieving the appropriate data, which means that the query identities must be transformed into human-readable language, providing data and metadata to support the language, and providing machine-actionable landing pages to enable the accessing of metadata and data via query re-execution (Rauber *et al.* 2015). The last step is making modifications to the data infrastructure, which includes technology migration for new data representation, and migration verification, which includes ensuring that queries can be re-executed correctly for the verification of successful data and query migration (Rauber *et al.,* 2015).

If the data citation is implemented correctly, then managing access to data will become more efficient. is increases the importance of successful data citation in order to manage access to data successfully.

#### 3.2.2.3 Quality control

When considering the quality control of big data, it may not be clarified easily due to big data being a relatively new format, but Cai and Zhu (2015: 4) state that big data quality control does not depend only on the data's features, but also on the business environment in which the data is used. This also includes business processes and business users. Cai and Zhu (2015: 4) identified quality criteria for big data in terms of data quality dimensions.

The first data quality dimension is availability. This is defined as the level of convenience for the users of the big data to access the data and necessary related information. The availability dimension includes data quality elements such as accessibility to the data, timeliness of the data, and authorisation of the data (Cai & Zhu, 2015: 4).

The second dimension of big data quality is usability. This includes whether or not the data meet the needs of the users. Usability, thus, includes elements such as documentation of the data, credibility of the data, and the data's metadata (Cai & Zhu, 2015: 4).

The third quality dimension of big data is reliability. Reliability refers to whether or not the data can be trusted by the user to make further official use of. This dimension includes elements such as accuracy of the data, integrity of the data, consistency mechanisms of the data, completeness of the data, and auditability of the data (Cai & Zhu, 2015: 4).

The fourth big data quality dimension is relevance, which includes the degree of correlation between data and content, and the data users' expectations of the given data. Data relevance includes the element of adaptability of the data to the specific need of the user (Cai & Zhu, 2015: 4).

The final quality dimension of big data is presentation quality. This includes the valid description method for the given data, which then allows for a full understanding of the data by users (Cai & Zhu, 2015: 4). The elements of the presentation quality dimension are the readability of the data by the user and the appropriate structure of the data (Cai & Zhu, 2015: 4).

In terms of quality assessment of the big data, Cai and Zhu (2015: 57) devised a quality assessment process to be followed. This assessment is illustrated in Figure 3.8, which begins with determining the goals of the data collection, to the completion of the output results of the big data.

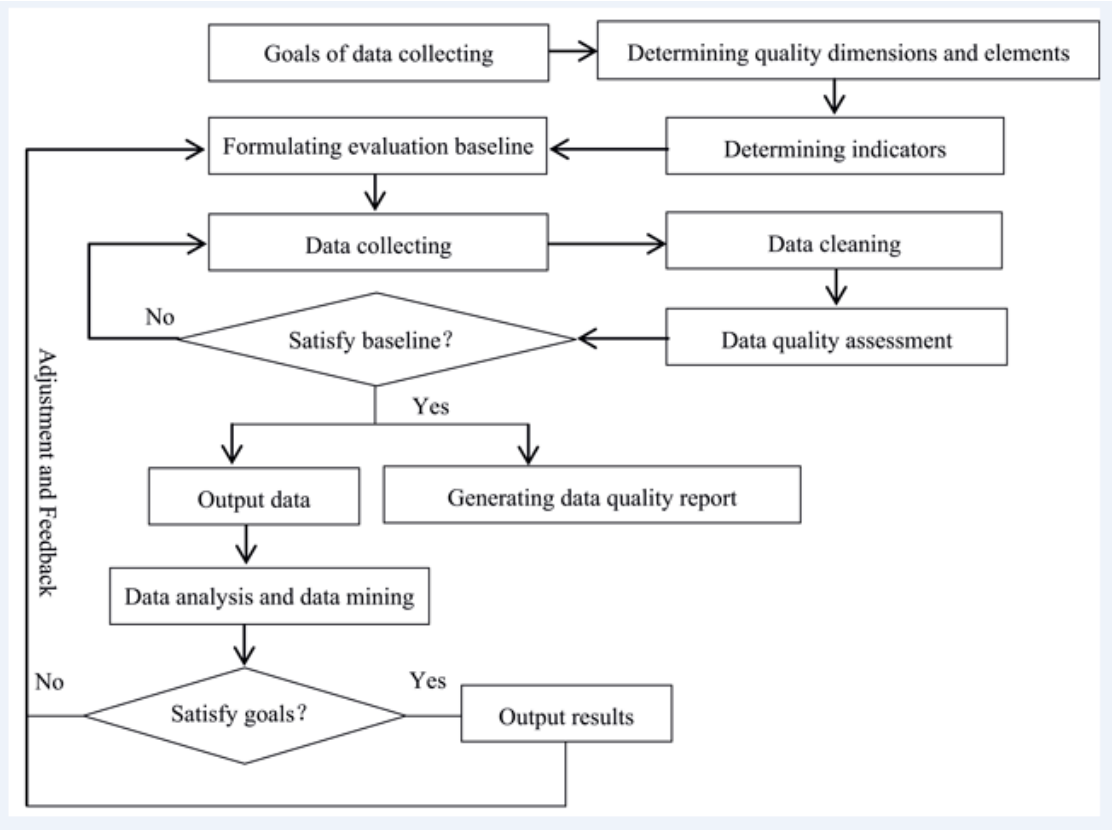


Figure 3.1: Quality assessment process for big data (Cai and Zhu, 2015: 7)

Figure 3.1 illustrates a quality assessment process for big data. This figure will assist with drawing appropriate conclusions for big data processes (Cai & Zhu, 2015: 7). The figure assists the big data user with determining the goals of the data collection, the selection of indicators which include data compliance, data collection and data cleaning, and data quality and mentoring phases (Cai & Zhu, 2015: 8). After the quality assessment and monitoring, an assessment takes place to track whether or not the data comply with the baseline evaluation made in the determination of indicators phase (Cai & Zhu, 2015: 8). If the data comply with the baseline evaluation, further assessment can continue with a follow-up data analysis and a data quality report. This is a useful process to determine data quality assurance, as it can be adjusted according to the needs of the environment.

#### 3.2.2.4 Applying FAIR principles

A primary responsibility of a data steward is to ensure that all data (also big data) under his or her stewardship adheres to the "FAIR" principles. These are a set of guiding principles which assist with making data **f**indable, **a**ccessible, **i**nteroperable and **r**eusable (Wilkinson *et al.* 2016: 4). Furthermore, the FAIR principles describe specific considerations for the current data environments, which support both manual and automated deposition, exploration, sharing and reuse (Wilkinson *et al.* 2016: 4).

The FAIR principles are iterative and concise, domain-independent, high-level principles that can be applied to a range of data outputs and environments, whereas, the term (meta)data offers the opportunity for the same principle to be applied to either the metadata or the data themselves (Wilkinson *et al.,* 2016: 4). FAIR is applicable to both humans and machines as it assists with data and (meta)data being machine-readable, which supports new knowledge discovery and innovations of multiple datasets (Force 11, 2016b). The FAIR principles are related to one another but are also independent and separable. Each of these principles contain and define the characteristics needed for third party discovery and reusability of the data, meaning that the barrier-to-entry for data producers and data stewards who intend making their data holdings according to the FAIR principles, are as low as possible (Wilkinson *et al.,* 2016: 4). More detail, about each of the principles, is provided below.

The findable principle is defined as the data object being uniquely and consistently identifiable when necessary. The data object should also be re-findable at any time, meaning that persistence is necessary with the data object's (meta)data (Force 11, 2016; Wilkinson *et al.* 2016: 4). The findable principle includes the data objects being machine-readable, which means it must contain appropriate (meta)data to be differentiated from other data objects (Force 11, 2016; Wilkinson *et al*. 2016: 4).

The accessible principle is defined as the (meta)data being retrievable by the identifier using a standardised communications protocol. This protocol should be free and universally implemented. The protocol should also allow for authentication and authorisation where necessary, and the metadata should be accessible even when the actual data is no longer available (Wilkinson *et al.* 2016: 4).

The third principle, namely, the interoperable principle, is defined as when the (meta)data used is in a formal, accessible, shared, and broadly applicable language for knowledge representation. The (meta)data use specific vocabulary and follow the FAIR principles. The (meta)data include qualified and legitimate references to other (meta)data (Force 11, 2016; Wilkinson *et al.* 2016: 4).

The last of the FAIR guiding principles is the reusable principle, which is defined as the (meta)data being described richly with a variety of accurate and relevant attributes, the (meta)data being released with a clear and accessible data usage licence, the (meta)data being associated with detailed sources, and the (meta)data meeting domain-relevant community standards (Wilkinson *et al.,* 2016: 4).

The FAIR principles are broad guiding principles for everyone to keep in mind when thinking about the data for which they are responsible. Only when the FAIR principles are applied, does the data become of value as a secondary resource. It is important to note that good data stewardship and the application of the FAIR guiding principles are interdependent and can only complement one another when applied meticulously.

Stewards using FAIR principles as guidelines will ensure that data assets reach their maximum potential, that the research at hand increases its visibility and citations, that reproducibility and reliability of the research is improved, and that the data and research remain in line with international standards and approaches (ANDS, 2017). Furthermore, the FAIR principles will ensure that the dataset attracts new partnerships with prospective researchers and relevant parties. The principles will enable new research to be discovered, will assist with the identification of new research approaches and tools, and will assist in achieving maximum impact for the research (ANDS, 2017).

#### 3.2.2.5 Ownership

Now that the roles have been clarified, the ownership and authority roles and responsibilities of the data steward can be discussed. The first ownership responsibility is shaping the data policy that should state clearly who will and who will not have access to institutional data. The data steward should also determine who approves or declines requests for access to data (Brubaker *et al.,* 2014: 4161). The procedures for requesting access to data should be aligned

with the policy (Brubaker *et al.,* 2014: 4161). The data steward should not only develop access procedures and processes: all the applicable documentation that guide the institution in the creation, collection and consumption of the data being stewarded must be developed (in collaboration, of course, with all the relevant role players). Lastly, the data steward has a responsibility to collaborate with other data stewards and to establish data standards where these may not exist within the research domains (Brubaker *et al.,* 2014: 4161). These responsibilities have to be carried out successfully to ensure the efficient stewardship of data as an intellectual asset of the organisation.

### 3.2.3 Challenges of big data stewardship

The characteristics of big data are considered to identify the challenges of big data. In this regard, Jagadish *et al.* (2014: 90) identify the first challenge of big data as heterogeneity, meaning that the data consumed is diverse to the extent that it is no longer possible for a human to understand and interpret data without computers. Computers, however, do not have human understanding, meaning that poor judgement by the computer with regard to the heterogeneous data, creating a rather large amount of data, which include both relevant and non-relevant data to be interpreted by the human being (Jagadish *et al.,* 2014: 90).

The second challenge of big data is inconsistency and incompleteness, as big data include data from diverse sources, with a varying amount of reliability and relevance. With this inconsistency in sources, uncertainty, errors, and missing data need to be managed and identified appropriately by the data steward or manager to avoid the inconsistency which these data may create, which in the end will also cause inconsistency in actionable data provided for the organisation or data user (Jagadish *et al*., 2014: 91).

The third challenge with regard to big data is scale, this is perhaps one of the most obvious challenges, as big data are large and ever-increasing in size, making it difficult to manage. The problem is that the scale of big data is increasing faster than CPU (central processing unit) speeds of computers, causing a lag in the processing of big data (Jagadish *et al*., 2014: 92).

The fourth notable challenge with big data is timeliness. Real-time techniques are needed to keep up-to-date with the ever-growing amounts of data so that data overload does not occur. This also means that the data managers and infrastructure used to manage the data need to keep pace constantly with the data (Jagadish *et al.*, 2014: 92).

The fifth challenge of big data is privacy and ownership, the privacy of the data of course of concern because of the personal data being accessed and exploited. This also affects the ownership of big data. It needs to be pointed out that big data are easily accessible, so anyone can access the data and claim that they are theirs. The management of the privacy and ownership of big data is both technical and sociological and needs to be addressed from both perspectives so that the benefits of big data are not lost (Jagadish *et al*., 2014: 92).

The last challenge pertaining to big data is visualisation and collaboration from the human perspective, meaning that for the big data to be useful, it must be on a scale which is absorbable by humans. The big data that are understood by the system are not always understood and absorbed appropriately by humans, which is the goal, after all. This means that the management process of big data needs to be of a high standard and must be accurate for humans to benefit from them as well, and not only the system (Jagadish *et al.,* 2014: 93). A solution can be found for each of the problems mentioned above, so that big data can be a positive opportunity for organisations and people, and not a problem.

*Table 3.1: Roles and responsibilities summarised*

| **Roles of the data steward** | **Responsibilities** |
|---|---|
| Champion | Facilitating and managing the appropriate and efficient use of the big data asset<br>Ensuring that data is FAIR (findable, accessible, interoperable and reusable) |
| Collaborator | Working collaboratively in teams<br>Building relationships with other data stewards<br>Collaborating with other data stewards to establish data standards where these may not exist |
| Technical expert | Interpreting heterogeneity in big data<br>Managing inconsistency and incompleteness in the data set<br>Being able to address issues of scale in the big data set<br>Making use of real-time techniques to keep up to date with the ever-growing amounts of data so that data overload does not occur<br>Managing the privacy and ownership of big data without compromising the value of the data<br>Finding ways to visualise data so that the data is useful from the human perspective<br>Organising data (skills to illustrate understanding of the data and the purpose of the data) |
| Knowledgeable expert | Maintaining expert level knowledge about:<br>• big data life cycles<br>• funder requirements<br>• the value of big data<br>• discipline knowledge, including discipline-specific methodologies<br>• Safety, security licensing and copyright of data<br>Developing a good understanding of research ethics |

| Roles of the data steward | Responsibilities |
|---|---|
| | Understand the importance of data as a secondary resource |
| Trainer | Facilitation, data presentation, understanding of learning styles, and working with different types of people<br>Training colleagues and researchers on the following topics:<br>• naming conventions,<br>• versioning,<br>• labelling<br>• writing data management plans<br>• writing data documentation<br>• allocating metadata |
| Quality assurer | Developing and promoting data quality standards<br>Ensuring that the data is associated with the correct metadata<br>Ensuring availability of the data and ensuring accessibility to the data<br>Considering the timeliness of the data<br>Checking the authorisation of the data<br>Ensuring **usability** (which means documentation of the data, credibility of the data, and the data's metadata)<br>Taking responsibility for the **reliability** of (trust in) of the data (which includes accuracy of the data, integrity of the data, consistency mechanisms of the data, completeness of the data, and auditability of the data)<br>Taking care of **relevance** issues (which includes the degree of correlation between data and content, and the data users' expectations of the given data, being |

| Roles of the data steward | Responsibilities |
|---|---|
| | able to advise on the adaptability of the data to the specific need of the user<br>Providing a valid description method for the given data, (which then allows for the full understanding of the data by users) |
| Gatekeeper | Facilitating and managing the appropriate and efficient use of the big data<br>Managing access to data<br>Establishing and maintaining repositories<br>Controlling data security<br>Ensuring the persistent identification of data sets<br>Measuring the impact of data sets |
| Access provider | Providing machine actionable landing pages enable the accessing of metadata and data via query re-execution<br>Considering query uniqueness<br>Ensuring that the sorting of data sets is unambiguous and reproducible<br>Setting up a standardised result set verification system,<br>Timestamping queries (when they are resolved compared to when they were made)<br>Query identification<br>Storing query metadata<br>Creating automated citation texts |

| Roles of the data steward | Responsibilities |
|---|---|
| Ownership arbitrator | Shaping the data policy<br>Developing access procedures and processes<br>Develop documentation that guide the institution in the creation, collection and consumption of the data<br>Licensing data<br>Guiding the proper citation of data |
| Data migrator | Making modifications to the data infrastructure (which includes technology migration for new data representation)<br>Migration verification (which includes ensuring that queries can be re-executed correctly for the verification of successful data and query migration) |
| Preservation manager | Archiving data<br>Managing data storage |
| Project manager | Project planning and management<br>Communicating new and changed business requirements to individuals affected<br>Managing time and deadlines<br>Working independently<br>Developing budgets |

Source: Brubaker, et al. 2014; Jagadish, et al., 2014; Cai and Zhu, (2015); Rauber *et al.* (2015); Malinowski, 2016; Segment, (2018) Lyon & Mattern, 2016; Seiner (2013); Brown, et al. 2015; Lyon, Matter, & Brenner (2016: 3-4), DataONE, 2018; RDA, 2018; WDS, 2018; IFLA, 2018; ANDS, 2018

The responsibilities identified (as they are linked to the roles) assume that the necessary skills and competencies, to be able to perform the responsibilities, are in place so that the desired outcome, well-managed data, is accomplished.

## 3.3 Skills, competencies, and outcomes

In order to understand the content of big data skills and the roles the big data steward plays in any organisation, it is important to define what a skill is, what a competency is, what a learning outcome is and how these terms are used in this context. This study focusses on specific skills for big data stewards, as these skills will fulfil the competencies, in the end, so that the roles required of a big data steward could be played successfully. It is important to recognise the difference between skill and competency to understand why skills are more applicable to this research. Several skills together may form a competency, or some skills address more than one competency. Outcomes are also defined, as in this study an outcome is the desired result, which is hoped to be obtained by the future learning modules. The desired outcomes will guide a relevant curriculum to prepare big data stewards for their future role. The research objective will, therefore, be achieved through the process of identifying relevant skills for big data stewards.

### 3.3.1 Skills defined

As already defined in the terms and definitions section, to recap and elaborate, a skill can be defined as the ability to do something well, this being the expertise that one has gained to perform a specific task successfully (Mamabolo *et al.,* 2017). Furthermore, a skill can be defined as the ability or the capacity that an individual has or has attained to fulfil a specific function. Skills in this sense are acquired deliberately and systematically (Mamabolo *et al.,* 2017).

### 3.3.2 Competencies defined

As already defined in the terms and definitions section, to recap, competencies usually are related to either a role to perform or a task to be completed within a given context or situation. A competency can be defined as a collection of related abilities, knowledge, and skills which assist individuals to fulfil their role or responsibility within their given environment (Fukada, 2018: 1-7). Furthermore, a competency can be defined as an individual performing a specific role based on their attained (1) skills, (2) past experiences, and (3) knowledge which they then apply to fulfil their role within their organisation (Fukada, 2018: 1-7). A competency can also be viewed as a responsibility which may occur at any stage of an individual's life,

inside or outside their organisation, where they need to act on or be responsible for a situation. This is where the individual's competencies come into play and the individual applies their **knowledge, skills, and past experiences** to be competent in a given situation.

### 3.3.3 Outcome defined

An outcome can simply be defined as the way that a situation turns out. This can be the way a situation ends up consequently, or because of specific actions throughout a situation (Zeppieri & George, 2017: 29). Furthermore, an outcome is the end of a sequence which has an input, a process, and an outcome. An outcome is often something that the process is wanting to achieve, whether it happens or not. Each situation or process ought to have a desired outcome at the end. The purpose of the process is to achieve the desired outcome (Zeppieri & George, 2014: 29).

Learning outcomes are important to take into consideration when designing curricula. A formal teaching module should include skill and knowledge outcomes. This research will establish how other academic institutions address these, and whether and how the experience component of competency development is being attended to.

## 3.4 Identified knowledge components required for data stewardship

A competent data steward does not only need specific skills to perform data stewardship responsibilities. The steward also needs a strong theoretical understanding of the following: big data life cycles, funder requirements, and the value of big data. It is likely that the steward would need discipline knowledge, including discipline-specific methodologies. Safety, security licensing and copyright form the next knowledge cluster, while the steward would also need to understand research ethics and the importance of data as a secondary resource. Each of these is described in more detail below.

### 3.4.1 Understanding life cycles with a focus on the big data life cycle

A steward needs to understand the data life cycle conceptually. This requirement was identified by authors such as Lee, Tibbo and Schaefer (2007), Lyon (2012: 129-130), Kim, Moen and Warga (2012: 69), Seiner (2013: 15-16), Jones, Pryor and Whyte (2013: 2), the DCC

(DCC, 2018), the Research Data Alliance (RDA), and the International Federation of Library Associations and Institutions (IFLA, 2018).

The data life cycle relates to all the sequential phases, which data moves through from collection to purging or destruction. According to the authors above, this skill requires that the steward has a solid understanding of the process of producing, creating, updating, deleting, retiring, and archiving the data which he or she is to manage (Seiner, 2013: 15-16).

### 3.4.2 Understanding funder requirements

Authors such as Lyon (2012: 129-130) and Lyon, Mattern and Brenner (2016: 3-4) identified knowledge about data funder requirements as essential. Funder-policy requirements and making research data management plans to address funder requirements for specific data projects (Lyon, 2012: 129-130; Lyon, Matter, & Brenner (2016: 3-4) are core knowledge items for a steward. Depending on the funder requirements, the steward may also need to know about discipline storage facilities and may need to understand how to collaborate with disciplinary, national and international data centres.

### 3.4.3 Understanding the value of [big] data as an asset

Understanding the value of data was identified by authors such as Lee, Tibbo and Schaefer (2007), Seiner (2013: 15-16), and the DCC life cycle model (DCC, 2018). Understanding the value of data would include a basic understanding of what data is, how it exists and the role it plays in its environment (Lee, Tibbo & Schaefer, 2007; Seiner, 2013: 15-16). It is also important to be able to recognise relevant and valuable data when presented with it (DCC, 2018).

### 3.4.4 Discipline-specific knowledge

Knowledge about the specific research discipline was identified by authors such as Lee, Tibbo and Schaefer (2007), Molloy and Snow (2012: 106-107). Burton, Lyon, Erdmann and Tijerina, (2017: 19), and DataONE (2018).

This knowledge is important because each discipline also has its own data requirements. Molloy and Snow (2012: 106-107) and Burton *et al.* (2017: 19) identified the need for disciplinary skills, saying that knowledge of what constitutes research data across a variety of

disciplines is necessary for effective stewardship. Lee, Tibbo and Schaefer (2007) further identified the need for understanding the professional and organisational environment in which the discipline data have to be curated.

### 3.4.5 Understanding discipline-specific research methodologies

Understanding a discipline is not sufficient. When working in different data and research contexts, the steward must also understand how research methodologies differ between disciplines. Molloy and Snow (2012: 106-107) state the importance of a discipline-specific understanding and the content which it contains, as this will assist the data steward in understanding how to deal with data from different disciplines. The foundational activities of dealing with the data may be similar, but because of the difference in the discipline, the data may be of a different form and will require a diverse knowledge base and understanding by a data steward to be utilised efficiently.

### 3.4.6 Safety and security

Knowledge of the safety and security issues associated with data is also important. Florentine (2017) has identified eight security skills which often have to be managed by the IT team. These include (1) security tools expertise (which involves knowing the tools and knowing the security functions for which the tools are needed); (2) security analysis (which involves the placement of IT tools in the organisation's security strategy, identifying system attacks and taking action to minimise attacks in future); (3) project management (which includes managing security projects for the IT team and ensuring that all security applications and tools are in place and doing their jobs); (4) incident response (which involves the securing of IT systems and the assistance of responding to system threats to the data quickly and efficiently, thus requiring the skills and knowledge to do so); (5) automation skills (which are systems set up to monitor systems and data for any threats automatically, data analytics skills, which involve the tracking of threats and potential attacks, and scripting skills, which involve getting all parts of the system to work together to protect the data efficiently); and (6) soft skills (which involve communication, collaboration, and teamwork to ensure that all tools and systems are in place to protect the data) (Florentine, 2017). It is important for the data steward to have knowledge of these skills to ensure the efficient protection and security of the data.

### 3.4.7 Licensing and copyright protection

It is also important for the data steward to have the necessary knowledge to protect the data with respect to laws, licensing and copyright. All legal rights regarding data protection should be enforced. The protection of data and the knowledge that data stewards should have regarding these laws are discussed in more detail in section 2.5.

### 3.4.8 Understanding of research ethics, specifically for data collection and manipulation

Clear knowledge and understanding of data ethics were identified by authors such as Lyon, Mattern and Brenner (2016: 3-4), Seiner (2013: 15-16), Lee, Tibbo and Schaefer (2007), and Jones, Pryor and Whyte (2013: 2). Data ethics is an important part of any work conducted with data. The integrity of data usage, the integrity and quality of data definition, and the integrity and quality of the data created or updated in the data process are all recognised parts of data ethics (Seiner, 2013: 15-16). Lee, Tibbo and Schaefer (2007) and Jones, Pryor and Whyte (2013: p2) identify the importance of mandates, values, principles and best practice throughout the process of working with the data. The DCC (2018) identifies knowledge of ethical practices as a part of a data curation checklist. De Sherbinin, Faustman and Edmunds (2018) and DataFirst's Siljeur (2018) agree with the sentiment.

### 3.4.9 Understanding data as a secondary research source

Ajayi (2017: 2) defines secondary data as that which is collected and produced by other users, while primary data are collected for the first time by the researcher. As the need to use data as a secondary source increases, it is important for the data steward to understand the secondary use of data. Secondary use relates to the reanalysis and reinterpretation of the primary data, but more often, it is about using data collected with one objective in mind to address a new objective. It could also be that a number of smaller datasets are combined to form a single big dataset.

## 3.5 Identified technical skills required for successful data stewardship

As this research is focussed on identifying a suitable training programme for big data stewards, it is necessary to list the required skills to steward data effectively. These skills have been ascertained in the literature and are discussed further in the sections below. Although

Digicurve (2013) has identified specific skills and competency levels of digital curation professionals, for the case of this study, Digicurve's skills will not be used as the skills identified by Digicurve are aimed more at the professional environment whereas this research is more aimed at the academic environment.

### 3.5.1 Writing data management plans

Data management planning skills are general skills pertaining to how data are managed successfully for their given purpose. This includes the preparation of data management plans when embarking on a new data project, and the cost of data management activities (Jones, Pryor & Whyte, 2013: 2). Data management planning skills were identified by authors such as Jones, Pryor and Whyte (2013: p2), the DCC life cycle model (DCC, 2018), the Australian National Data Service (ANDS, 2018), De Sherbinin, Faustman and Edmunds (WDS, 2018), and the Research Data Alliance (RDA, 2018).

The DCC (2018) also mentions the importance of data management skills, which include all pre-planning, during, and post-planning of the data project.

The writing of data plans is also part of data management as it entails all the components of a data project (De Sherbinin, Faustman, & Edmunds, 2018), and successful data management practice, as stated by the Research Data Alliance (RDA, 2018).

### 3.5.2 Administrative data documentation

Data administrative skills include all the general administrative skills that accompany the successful management of data, such as written communication and documentation data skills (Lyon & Mattern, 2016: 4), specific data functions and skills to assist with digital curation of data (Lee, Tibbo. & Schaefer, 2007), and advice and skills for data documentation on relevant systems (Jones, Pryor, & Whyte, 2013: 2).

Data administrative skills were identified by authors such as Lee, Tibbo and Schaefer (2007), Jones, Pryor and Whyte (2013: 2), Lyon and Mattern (2016: p4), the DCC life cycle model (DCC, 2018), De Sherbinin, Faustman and Edmunds (WDS, 2018), and DataONE (DataONE, 2018).

### 3.5.3 Developing data policy and procedural documentation

Data policy development was identified as a skill needed for successful research data management (RDM) (Brown *et al.*, 2015: 7). Data policy development skills were identified by authors such as Brown *et al*. (2015: 7), and the Research Data Alliance (RDA, 2018).

The Research Data Alliance also recognised data policy development as a learning resource for data-related projects (RDA, 2018).

### 3.5.4 Data appraisal (evaluation and assessment of relevant data)

Data appraisal skills were identified by authors such as Molloy and Snow (2012: 106-107), Lyon (2012: 129-130), and the DCC life cycle model (DCC, 2018).

It is important to note that data appraisal skills can further be defined as the evaluation and assessment of relevant data, which exist within an organisation or the given environment. This data could be appraised for a variety of reasons, but the main reason is to ascertain if the data are serving their purpose within the organisation (Molloy & Snow, 2012: 106-107). Furthermore, data appraisal skills can be acquired by understanding the data at hand and their purpose, by means of gathering and evaluating the relevant data (Molloy & Snow, 2012: 106-107). The skills can also be acquired by understanding the data in terms of which data to keep and which to discard (Lyon, 2012: 129-130).

### 3.5.5 Licensing of data

Data licensing skills were identified by authors such as Lyon (2012: 129-130), Jones, Pryor and Whyte (2013: 2), Brown, Bruce, and Kernohan. (2015: 7), De Sherbinin, Faustman and Edmunds (WDS, 2018), and the Research Data Alliance (RDA, 2018).

Brown *et al.* (2015: p7) recognise advocacy as a part of data licensing skill control as it pertains to the legal work of data. Regulatory compliance also plays its part in data licensing, as it is adherence to laws, regulations and guidelines (Jones, Pryor, & Whyte, 2013: p2). Research data management licensing includes guidance with queries regarding data licensing, which also involves the legal and ethical issues associated with datasets (Lyon, 2012: 129-130). The WDS considers standardisation, licences and intellectual property rights as part of recognised

RDM skills (De Sherbinin, Faustman, & Edmunds, 2018). Finally, the Research Data Alliance recognises data licensing and privacy as data skills (RDA, 2018).

### 3.5.6 Data archiving

Data archiving skills were identified by authors such as Molloy and Snow (2012: 106-107), Jones, Pryor, and Whyte (2013: 2), Brown *et al.* (2015: 7), the DCC life cycle model (DCC, 2018), the International Federation of Library Associations and Institutions (IFLA, 2018), and Siljeur (DataFIRST, 2018).

Data archiving skills are important, as recognised by several authors. Brown *et al.* (2015: 7) recognise the RDM skills of data archiving and preservation of the data. Molloy and Snow (2012: 106-107) also report an awareness of data preservation and curation options. Jones, Pryor and Whyte (2013: 2) have identified the post-data project skill of archiving data, and necessary retention periods of data as identified by IFLA (2018). Finally, Siljeur (2018) from DataFIRST, also regards data archiving as skills essential for the improvement of data products in the data process.

### 3.5.7 Allocating metadata

Metadata skills were identified by authors such as Lyon (2012: 129-130), Kim, Moen and Warga (2013: 69), DataONE (DataONE, 2018), Brown *et al.* (2015: 7), Siljeur (DataFIRST, 2018), the Research Data Alliance (RDA, 2018), De Sherbinin, Faustman and Edmunds (WDS, 2018), the Australian National Data Service (ANDS, 2018), and the DCC life cycle model (DCC, 2018).

Metadata entail a set of data which give information about and meaning to another set of data. Understanding metadata is important when working with any set of data, and within any data process, hence, this term is mentioned by numerous authors, as noted above. It includes metadata cataloguing, metadata guidance, technical expertise for structured data description with metadata standards, general metadata skills, metadata formats, usage and data discovery, metadata generators, metadata creation, metadata management, and metadata development.

### 3.5.8 Developing data citations

Data citation skills mentioned by Lyon (2012: 129-130), and the Research Data Alliance (RDS, 2018) and are recognised as supplying guidance and links to third-party services (Lyon, 2012: 129-130), and research data citation and the citing of data as recognised by the RDA (, 2018).

### 3.5.9 Project management

Project management skills are identified by authors such as Seiner (2013: 15-16), Kim, Moen and Warga (2013: 69), Brown *et al.* (2015: 7), Lyon and Mattern (2016: 4), the Research Data Alliance (RDA, 2018), and DataONE (2018).

Project management skills may include diverse skills in the data project and process, and business analysis, as stated by Brown *et al*. (2015: 7). Seiner (2013:15-16) recognised a project management skill as communicating new and changed business requirements to individuals affected. Project planning and management were identified by Kim, Moen and Warga (2013: p69). DataONE (2018) identified the project director as having an important skill and role to be developed in data management processes and projects. The inclusion of community of practice was identified by RDA (RDA, 2018). Lastly, the ability to work collaboratively in teams or with clients also forms part of a project management skill within the data process (Lyon, Mattern & Brenner, 2016: 4).

### 3.5.10 Data usage skills

Data usage skills are mentioned by authors such as Kim, Moen, & Warga (2013: 69), the DCC life cycle model (DCC, 2018), and DataONE (2018).

Data usage skills are recognised as a form of data collection by Kim, Moen and Warga (2013: 69), and as data access and usage, data collection, and data recognition by the DCC life cycle model (DCC, 2018). DataONE also refers to data usage under a data collection spectrum, and the use of data by means of finding data for reuse, citing data, data analysis tools, support services, and data literacy (ANDS, 2018).

### 3.5.11 Research skills

Data research skills are recognised by authors such as Jones, Pryor and Whyte (2013: 2), Lyon and Mattern (2016: 4) and Lyon, Mattern and Brenner (2016: 3-4).

Data research skills are defined by the above authors as the skills necessary to research the relevant data and to recognise the relevant data when needed, for whichever purpose. To understand data research skills, Lyon and Mattern (2016: 4) state that one needs practical research skills and experiential research skills as gained by the researcher. Lyon, Mattern and Brenner (2016: 3-4) identify the necessary data research skills such as understanding research workflows in specific disciplines served, meaning that research skills should be applied to different disciplines no matter how different the workflow may be. Lastly, to understand data research skills it is also important to make research data visible within any data process, meaning that data research skills are necessary to make the product of the research data visible (Jones, Pryor, & Whyte, 2013: 2).

### 3.5.12 Working with data skills and data formatting skills

A host of practical and technical data skills were identified by authors such as Barth, Bean, & Davenport (2012: 22-24), Lyon (2012: p129-130), Jones, Pryor and Whyte (2013: 2), Kim, Moen and Warga (2013: 69), Lyon and Mattern (2016: p4), Burton *et al*. (2017: 19), the International Federation of Library Associations and Institutions (IFLA, 2018), and DataONE (DataONE, 2018).

Practical and technical skills identified by various authors entail diverse skills, such as computational skills, quantitative skills, and technical skills as noted by Burton *et al*. (2017: 19). Understanding software is recognised by Kim, Moen and Warga (2013: 69), while the use of inline tools is identified by Jones, Pryor and Whyte (2013: 2). Data analytics, infrastructure, framework and application, and data mining and analysis are all the skills recognised by the IFLA (2018). Barth, Bean and Davenport (2012: 22-24) identify creative IT skills, programming skills, mathematical skills, statistical skills, and analytic skills for a big data skillset. DataONE recognise the data analyser as one who plays an important role in data management. RDM planning, including exploring future data infrastructure demands is identified by Lyon (2012:

129-130). Lastly, Lyon, Mattern & Brenner (2016: 4) state the organisational and analytical skills needed for the data process and for the data curation role.

Data formatting skills are indicated by authors such as Molloy and Snow (2012: 106-107), Lyon, Mattern and Brenner (2016: 3-4), Siljeur (DataFIRST, 2018), the Research Data Alliance (RDA, 2018), De Sherbinin, Faustman and Edmunds (WDS, 2018), and Jones, Pryor and Whyte (2013: 2).

### 3.5.13 Managing data storage

Data storage skills are mentioned by authors such as Lyon (2012: 12-130), Jones, Pryor and Whyte (2013: 2), DataONE (2018), the Research Data Alliance (RDA, 2018), De Sherbinin, Faustman and Edmunds (WDS, 2018), and the DCC life cycle model (DCC, 2018).

Data storage skills, as identified by the above authors play an important role in any data process. Jones, Pryor and Whyte (2013: 2) recognise that throughout the data process, data storage facilities are necessary. Lyon (2012: 129-130) also regard RDM storage as necessary to ensure clarity and relevance of local data storage guidelines and infrastructure provision. The DCC life cycle model (DCC, 2018) have also identified storage skills as necessary during the curation process. The WDS regard data organisation and storage as necessary big data skills (De Sherbinin, Faustman, & Edmunds, 2018). In turn, DataONE has identified the computing backup and storage roles and skills needed for data management. Furthermore, secure storage for research data is mentioned by the Research Data Alliance (RDA, 2018).

### 3.5.14 Establishing and maintaining repositories

Data repository skills are noted by authors such as Lyon, Matter, & Brenner (2016: 3-4), DataONE (2018), and the Research Data Alliance (RDA, 2018).

Data repository skills include the understanding of a data repository and the role it plays in the data management process, and data curation process. Lyon, Mattern and Brenner (2016: 3-4) note that familiarity is needed with relevant disciplinary data repositories to enhance successful data management. Furthermore, DataONE identifies the data model or database designer as an important role and skill to have within the data management process. The Research Data Alliance (RDA, 2008) refers to the data cataloguing or data repository learning resource and an understanding of them.

### 3.5.15 Developing infrastructure

The maintenance of infrastructure for big data stewards would include the establishment and maintenance of appropriate repositories. This includes identifying an appropriate repository at the start of a project, after which the curator should ensure that the researchers involved, adhere to the requirements of the selected repository. This includes giving appropriate access to storage for repository content. Gibson (2013: 3-138) indicates eleven steps for the establishment of a repository. Even though all the steps will be mentioned, it is important to note that in the case of the big data stewards, special focus should be on steps one, three, four, six, eight, nine, and eleven.

The steps are as follows:

Step 1 – Create an open access and digital preservation plan. This would include writing an open access policy, an institutional digital preservation policy, and a plan for building capacity. These activities enhance the long-term system availability (sustainability) of the repository (Gibson, 2013: 3; Marsh, Wackerman & Stubbs, 2017: 3-5).

Step 2 – This step includes the persistence of identifiers (DOIs/Handles/URL's) which are extremely important for access to the resource. It is also essential for electronic citation monitoring in the longer term, and for the marketing of the repository (Gibson, 2013: 6; Marsh *et al.* 2017: 3-5).

Step 3 – This step involves the appointment of repository management personnel and assigning repository librarians and a repository manager to the project (which is where the curator role manifests itself) and contracting a system administrator and a website programmer (Gibson, 2013: 8). The administrator and website programmer are probably not full-time positions, but the skills are required on an ad hoc basis. The repository librarian and the repository manager roles could be fulfilled by one individual.

Step 4 – Acquiring the hardware: This step focused on the building of repositories and IT infrastructure. The curator is involved with budgeting for and the purchasing of the server hardware and other hardware needed and considering the planning for the server replacement when the warranty expires (Gibson, 2013: 10; Marsh *et al*. 2017: 3-5).

Step 5 – Commissioning the server and installation of the repository software. This step includes the acquisition of skills such as Linux, to install and maintain the Linux servers, and any other software skills that are applicable.

Step 6 –Repository system backup and monitoring. The step includes planning for disaster recovery and long-term system sustainability (Gibson, 2013: 63; Marsh *et al*., 2017: p3-5).

Step 7 –The repository launches and the registration with harvesters, including the planning of an official repository launch, and the registration of the repository with as many harvesters as possible to enhance the visibility of the repository (Gibson, 2013: 65; Marsh, Wackerman, & Stubbs 2017: 3-5).

Step 8 – This step includes the capturing of research records and the submission of research items. This kind of capturing of research will increase the usability of the repository and the impact it may have. Peer-reviewed articles, dissertations, and as many relevant pieces of research as possible should be ingested (Gibson, 2013: 71; Marsh *et al*., 2017: 3-5).

Step 9 – This step involves the repository self-help and news which includes the setup of the self-help wiki and platform for normal users of the repository, and the setting up of news blogs to keep all the users of the repository informed (Gibson, 2013: 74; Marsh *et al.* 2017: 3-5).

Step 10 – This step pertains to the engaging of research partners for the repository, this includes third parties and any researchers who may benefit from submitting into the repository, and the repository gaining a mutual benefit (Gibson, 2013: 75; Marsh *et al*., 2017: 3-5).

Step 11 – The last step involves the continuous improvement of the system, this will include constant analyses of the system to ensure that the repository is constantly up to date and meeting the needs of the users (Gibson, 2013: 8; Marsh *et al.,* 2017: 3-5).

With the above steps regarding how one would establish a repository, or infrastructure, a big data steward would be involved and would operate with the system accordingly to ensure that once the repository is created, that they are able to handle the data appropriately.

### 3.5.16 Data organisation skills

Data organisation skills were identified by authors such as De Sherbinin, Faustman, & Edmunds (WDS, 2018), and the International Federation of Library Associations and Institutions (IFLA, 2018).

Data organisation skills are also extremely important skills to illustrate understanding of the data and the purpose of the data. The WDS recognised the skill of data organisation, and IFLA recognised the need for the skill of organising and accessing data.

### 3.5.17 Data accessibility, dissemination and sharing

Data access skills have been identified by authors such as Jones, Pryor and Whyte (2013: 2), Siljeur (DataFIRST, 2018), the Research Data Alliance (RDA, 2018), the DCC life cycle model (DCC, 2018), and the International Federation of Library Associations and Institutions (IFLA, 2018).

Data access is an essential skill needed in any part of a data process or project. Accordingly, Jones, Pryor and Whyte (2013: 2) note that data access is a skill needed throughout the project as the recognition of data types, platforms and access to knowledge, and by the DCC life cycle model as data retrieval skills. IFLA also recognise the need for organising and accessing data, and that of supporting data discovery was identified by Siljeur (2018) from DataFIRST. Lastly, the Research Data Alliance (RDA, 2018) regard the learning resources of data access, discovery and sharing as important and necessary skills.

Data sharing skills were identified by authors such as the DCC life cycle model (DCC, 2018), Molloy and Snow (2012: 106-107), and Seiner (2013: 15-16).

Data sharing is also an important skill to have in any data project or process. It is described by Molloy and Snow (2012: 106-107) as the knowledge of data-sharing options, including which licence is right for the researcher's needs, and is regarded by Seiner (2013: 15-16) as the need for supporting and sharing data and knowledge with other data stewards. Finally, the DCC life cycle model also identifies data sharing as an integral part of any data management or curation process or project.

Data dissemination skills are noted by authors such as Siljeur (DataFIRST, 2018), the Australian National Data Service (ANDS, 2018), the DCC life cycle model (DCC, 2018), and Kim, Moen and Warga (2013: 69).

### 3.5.18 Training skills

Training can also be addressed as an important skill for a data steward. Training skills may involve data facilitation, data presentation, understanding of learning styles, and working with different types of people. Training skills may form part of the soft skills necessary as they involve personal skills as well as the skills necessary to adapt and work with different people and in different environments (ANDS, 2018).

### 3.5.19 Technical data skills

With regard to the technical skills required by the data steward, it is important to note that all the data stewards will acquire and need different skills according to the discipline in which they are involved. According to ANDS (2018), technical skills for data stewards may include programming for data-intensive research, environmental data knowledge, visualising environmental data, data management, interdisciplinary data exchange. Software development, data organisation, and data repositories.

## 3.6 Personal (soft) skills required for future work environments

Soft skills are those skills needed for the researcher or person involved to be able to interact successfully with others, as mentioned by Burton, Lyon, Erdmannand Tijerina (2017: 19). These skills are important in any project, as all data projects will include working and interacting with different individuals throughout the project.

### 3.6.1 Time management

Personal data skills would include all the skills needed to work with data successfully and to understand the data fully that are involved. These personal data skills include successful time management, the ability to work independently, and attention to detail and accuracy when working with data (Lyon, Mattern & Brenner, 2016: 4).

### 3.6.2 Independent worker

An important skill which the data steward needs is the ability to work independently. As mentioned, a data steward needs to be able to work with different people, but also needs to be able to work alone and to have all the necessary skills to fulfil the requirements of the data. This means that a steward's task will be to work on the data independently and complete a diverse number of tasks on the data, this is necessary if the steward is not in a team (Lyon, Mattern & Brenner, 2016: 4).

### 3.6.3 Paying attention to detail (accuracy)

Data stewards work with large amounts of data, meaning that they will also work with a great deal of detail. Stewards needs to be accurate and minimise mistakes when fulfilling tasks on the data and need to be able to pay attention to fine detail and understand the intricate tasks to be fulfilled. Data may vary and may change according to the environment in which they exists all the more reason for data stewards to have the relevant knowledge and skills with regard to the data in the particular environment, which will also lead to the stewards avoiding mistakes in the end due to inaccuracy or not paying attention to detail. The more data stewards know about the data, the more mistakes they can avoid (Lyon, Mattern & Brenner, 2016: 4).

### 3.6.4 Build relationships

As data stewards may find themselves working in different situations with different kinds of data, it is important for data stewards to be able to build relationships with other stewards or professionals in the field, and in their team. This is important because if the stewards need a task to be fulfilled, which they cannot, they can contact a professional with whom they have built a good relationship. Accordingly, it is important for stewards to build relationships with other stewards and other data professionals so that they can share information regarding the data and activities, and positive and negatives aspects regarding any happenings within the field (Lyon, Mattern & Brenner, 2016: 4).

### 3.6.5 Community-based data skills

Community based data skills were identified by authors such as the Research Data Alliance (RDA, 2018), and the DCC life cycle model (DCC, 2018).

In this regard, community-based skills include the skills and knowledge needed to interact with the community and to show how the data may fulfil the needs expressed by the community in which they exist. This is a skill which addresses the external client of the organisation, but the data curator needs to know the community needs and be able to assist with achieving a community-based goal with the data.

With regard to the identification of the above-mentioned skills, roles and responsibilities from the various sources and data initiatives, it is important to note that certain initiatives were not included because they do not address the issue of skills development. The first initiative (CODATA) was not included because their site was being updated, making the information inaccessible. The second and third initiatives (LIASA and FORCE11) were not included because they do not address this issue.

### 3.6.6 Developing budgets

As stated in step 4 of the development of infrastructure, the data steward is involved with the budgeting of the hardware and software needed for the repository to operate (Gibson, 2013: 10). The steward is therefore involved in the development of an appropriate budget to plan what is needed and how much money is needed to fulfil the needs of the repository and ultimately the needs of the user of the repository. Within the parameters of this budget, the steward should also consider warranty expiry dates and any legal issues which need to be included in the budget. In other words, the steward needs to ensure that all that is necessary should be included in the budget (Gibson, 2013)

## 3.7 Data competency matrix: knowledge, skills and experience required

Drawing on all the sources consulted above regarding the necessary skills, a data competency matrix was built to provide an overview of all the skills required, and from which source. The knowledge matrix below indicates all the knowledge-based skills which are needed for appropriate big data curation, as stated in the relevant sources.

*Table 3.2: Knowledge components required*

| | **Sources from which skills were identified** | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Knowledge component** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| Data life cycles (3.4.1) | | | | | | | **X** | **X** | **X** | **X** | **X** | **X** | | | | **X** | **X** | |
| Funder requirements (3.4.2) | | | | | | | | | | | | **X** | | **X** | | | | |
| Value of data assets (3.4.3) | | | | | | | X | | | | X | | | | | | X | |
| Discipline research requirements / methodologies (3.4.4&5) | | | | **X** | | **X** | | | **X** | | **X** | | **X** | **X** | **X** | | | |
| Safety & security (3.4.6) | | | | | | | | | | | | | | | | | | |

| | **Sources from which skills were identified** | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Knowledge component** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| Licensing and copyright (3.4.7) | | | | | | | | | | | | | | | | | | |
| Research ethics (3.4.8) | | | | | **X** | | **X** | | **X** | | **X** | | | | | | **X** | **X** |
| Secondary data use (3.4.9) | | | | | | | | | | | | | | | | | | |

**Legend**

| | | | | | |
|---|---|---|---|---|---|
| 1 | ANDS (2018) | 7 | DCC (Higgins: 136) | 13 | Lyon & Mattern (2016: 4) |
| 2 | Barth, Bean, & Davenport (2012: 22-24) | 8 | IFLA (2018) | 14 | Lyon, Mattern, & Brenner (2016: 3-4) |
| 3 | Brown, et al. (2015: 7) | 9 | Jones, Pryor, & Whyte (2013: 2) | 15 | Molloy & Snow (2012: 106-107) |
| 4 | Burton, et al. (2017: 19) | 10 | Kim, Moen, & Warga (2012: 69) | 16 | RDA (2018) |
| 5 | DataFIRST (Siljeur, 2018) | 11 | Lee, Tibbo, & Schaefer (2007) | 17 | Seiner (2013: p15-16) |
| 6 | DataONE (2018) | 12 | Lyon (2012: 129-130) | 18 | WDS (De Sherbinin, Faustmann, & Edmunds, 2018) |

The technical skills matrix below indicates all the technical skills needed for appropriate big data curation, as stated in the relevant sources.

*Table 3.3: Technical skills required*

| | Sources from which skills were identified | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Technical skills** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | | | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| Writing DMPs (3.5.1) | X | | | | | | X | | | | X | | | | | | | X | | X |
| Administrative skills (3.5.2) | | | | | | X | X | | | | X | | X | | X | | | X | | X |
| Developing policy & procedural documents (3.5.3) | | | X | | | | | | | | | | | | | | | X | | |
| Appraisal skills (3.5.4) | | | | | | | X | | | | | | | X | | | X | | | |
| Licensing of data skills (3.5.5) | | | X | | | | | | | | X | | | X | | | | X | | X |
| Archiving skills (3.5.6) | | | X | | X | | X | X | | | X | | | | | | X | | | |
| Metadata skills (3.5.7) | X | | X | | X | X | X | | | | | X | | X | | | | X | | X |
| Citation of data skills (3.5.8) | | | | | | | | | | | | | | X | | | | X | | |
| Project management skills (3.5.9) | | | X | | | X | | | | | | X | | | X | | | X | X | |
| Data usage skills (3.5.10) | X | | | | | X | X | | | | | X | | | | | | | | |

| | Sources from which skills were identified | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Technical skills** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | | | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| Data research skills (3.5.11) | | | | | | | | | | | | | | | | | | | | |
| Working with and formatting (3.5.12) | | | | | X | | | | | | X | | | | | X | X | X | | X |
| Managing data storage (3.5.13) | | | | | | X | X | | | | X | | | X | | | | X | | X |
| Repository skills (3.5.14) | | | | | | X | | | | | | | | | | X | | X | | |
| Organisation skills (3.5.15) | | | | | | | | X | | | | | | | | | | | | X |
| Access, sharing, dissemination skills (3.5.16) | X | | | | X | | X | X | | | X | X | | | | | | X | | |
| Training skills (3.5.17) | | | | | | | | | | | | | | | | | | | | |
| Practical and technical data skills (3.5.18) | | X | | X | | X | | X | | | X | X | | X | X | | | | | |

The soft skills matrix below indicates the soft skills needed for appropriate big data curation, as stated in the relevant sources.

*Table 3.4: Soft skills required*

| | **Sources from which skills were identified** | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Soft skills** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| Time management (3.6.1) | | | | | | | | | | | | | | | | | | |
| Independent worker (3.6.2) | | | | | | | | | | | | | | | | | | |
| Attention to detail (3.6.3) | | | **X** | | | | | | | | | | | | | | | |
| Building relationships (3.6.4) | | | | | | | | | | | | | | | | | | |
| Community-based data skills (3.6.5) | | | | | | | **X** | | | | | | | | | **X** | | |
| Developing budgets (3.6.6) | | | | | | | | | | | | | | | | | | |

## Legend

| | | | |
|---|---|---|---|
| 1 | ANDS (2018) | 11 | Lee, Tibbo, & Schaefer (2007) |
| 2 | Barth, Bean, & Davenport (2012: p22-24) | 12 | Lyon (2012: p129-130) |
| 3 | Brown, et al. (2015: 7) | 13 | Lyon & Mattern (2016: 4) |
| 4 | Burton, et al. (2017: 19) | 14 | Lyon, Mattern, & Brenner (2016: 3-4) |
| 5 | DataFIRST (Siljeur, 2018) | 15 | Molloy & Snow (2012: 9106-107) |
| 6 | DataONE (2018) | 16 | RDA (2018) |
| 7 | DCC (Higgins, 2018: p136) | 17 | Seiner (2013: 15-16) |
| 8 | IFLA (2018) | 18 | WDS (De Sherbinin, Faustmann, & Edmunds, 2018) |
| 9 | Jones, Pryor, & Whyte (2013: 2) | | |
| 10 | Kim, Moen, & Warga (2012: 69). | | |

### 3.8 Linking skills to roles and responsibilities

Now that the roles and responsibilities and the relevant skills have been identified, they can be linked to gain an understanding of what the data steward should be capable of doing. Please note that the skills below are the technical skills, soft skills, as well as knowledge components.

[Please refer to Table 3.1 in section 3.2.3 for the first two columns. Column 3 is debatable – each person would do this differently]

*Table 3.5: Skills linked to roles and responsibilities*

| **Roles of the data steward (refer to 3.2)** | **Responsibilities (refer to 3.2)** | **Skills necessary to address the responsibility identified (refer to 3.3 – 3.6)** |
|---|---|---|
| Champion | • Facilitating and managing the appropriate and efficient use of the big data asset.<br>• Ensuring that data is FAIR (findable, accessible, interoperable and reusable). | (S) Independent worker<br>(S) Community based data skills<br>(S) Building relationships |
| Collaborator | • Working collaboratively in teams.<br>• Building relationships with other data stewards.<br>• Collaborating with other data stewards to establish data standards where these may not exist. | (T) Administrative skills<br>(T) Project management skills<br>(S) Building relationships |
| Technical expert | • Interpreting heterogeneity in big data<br>• Managing inconsistency and incompleteness in the data set.<br>• Being able to address issues of scale in the big data set.<br>• Making use of real-time techniques to keep up to date with the ever-growing amounts of data so that data overload does not occur. | (T) Administrative skills<br>(T) Developing policy & procedural documents<br>(T) Appraisal skills<br>(T) Licensing of data skills<br>(T) Archiving skills<br>(T) Metadata skills<br>(T) Citation of data skills<br>(T) Project management skills<br>(T) Data usage skills |

| **Roles of the data steward (refer to 3.2)** | **Responsibilities (refer to 3.2)** | **Skills necessary to address the responsibility identified (refer to 3.3 – 3.6)** |
|---|---|---|
| | • Managing the privacy and ownership of big data without compromising the value of the data.<br>• Finding ways to visualise data so that the data is useful from the human perspective.<br>• Organising data (skills to illustrate understanding of the data and the purpose of the data). | (T) Data research skills<br>(T) Working with and formatting<br>(T) Managing data storage<br>(T) Repository skills<br>(T) Organisation skills<br>(T) Access, sharing, dissemination skills<br>(T) Practical and technical data skills |
| Knowledgeable expert | Maintaining expert level knowledge about:<br>• big data life cycles.<br>• funder requirements.<br>• the value of big data.<br>• discipline knowledge, including discipline-specific methodologies.<br>• Safety, security licensing and copyright of data<br>• Developing a good understanding of research ethics.<br>• Understand the importance of data as a secondary resource. | (K) Data lifecycles<br>(K) Research ethics<br>(K) Funder requirements<br>(K) Value of data<br>(K) Secondary data use<br>(S) Community based data skills |

| **Roles of the data steward (refer to 3.2)** | **Responsibilities (refer to 3.2)** | **Skills necessary to address the responsibility identified (refer to 3.3 – 3.6)** |
|---|---|---|
| Trainer | • Facilitation, data presentation, understanding of learning styles, and<br>• working with different types of people.<br>• Training colleagues and researchers on the following topics:<br>○ naming conventions.<br>○ versioning.<br>○ labelling.<br>○ writing. data management plans.<br>○ writing data documentation.<br>○ allocating metadata. | (S) Time management<br>(T) Writing DMPs<br>(T) Administrative skills<br>(T) Developing policy & procedural documents<br>(T) Appraisal skills<br>(T) Licensing of data skills<br>(T) Archiving skills<br>(T) Metadata skills<br>(T) Citation of data skills<br>(T) Project management skills<br>(T) Data usage skills<br>(T) Data research skills<br>(T) Working with and formatting<br>(T) Managing data storage<br>(T) Repository skills<br>(T) Organisation skills<br>(T) Access, sharing, dissemination skills<br>(T) Practical and technical data skills<br>(K) Data life cycles<br>(K) Funder requirements<br>(K) Value of data assets<br>(K) Discipline research requirements / methodologies<br>(K) Safety & security<br>(K) Licensing and copyright<br>(K) Research ethics<br>(K) Secondary data use |

| **Roles of the data steward (refer to 3.2)** | **Responsibilities (refer to 3.2)** | **Skills necessary to address the responsibility identified (refer to 3.3 – 3.6)** |
|---|---|---|
| Quality assurer | • Developing and promoting data quality standards.<br>• Ensuring that the data is associated with the correct metadata.<br>• Ensuring availability of the data and ensuring accessibility to the data,<br>• Considering the timeliness of the data,<br>• Checking the authorisation of the data<br>• Ensuring **usability** (which means documentation of the data, credibility of the data, and the data's metadata).<br>• Taking responsibility for the **reliability** of (trust in) of the data (which includes accuracy of the data, integrity of the data, consistency mechanisms of the data, completeness of the data, and auditability of the data).<br>• Taking care of **relevance** issues (which include the degree of correlation between data and content, and the data users' expectations of the given data, being able to advise on the adaptability of the data to the specific need of the user.<br>• Providing a valid description method for the given data, (which then allows for full understanding of the data by users). | (T) Appraisal skills<br>(T) Data research skills<br>(T) Access, sharing, dissemination skills<br>(T) Practical and technical data skills |

| **Roles of the data steward (refer to 3.2)** | **Responsibilities (refer to 3.2)** | **Skills necessary to address the responsibility identified (refer to 3.3 – 3.6)** |
|---|---|---|
| Gatekeeper | • Facilitating and managing the appropriate and efficient use of the big data.<br>• Managing access to data.<br>• Establishing and maintaining repositories.<br>• Controlling data security.<br>• Ensuring the persistent identification of data sets.<br>• Measuring the impact of data sets. | (T) Developing policy & procedural documents<br>(T) Appraisal skills<br>(T) Licensing of data skills<br>(T) Metadata skills<br>(T) Repository skills<br>(T) Access, sharing, dissemination skills |
| Access provider | • Providing machine actionable landing pages enable the accessing of metadata and data via query re-execution.<br>• Considering query uniqueness.<br>• Ensuring that sorting of data sets is unambiguous and reproducible.<br>• Setting up a standardised result set verification system.<br>• Timestamping queries (when they are resolved compared to when they were made).<br>• Query identification.<br>• Storing query metadata.<br>• Creating automated citation texts. | (S) Community based data skills<br>(T) Administrative skills<br>(T) Developing policy & procedural documents<br>(T) Licensing of data skills<br>(T) Metadata skills<br>(T) Citation of data skills<br>(T) Data usage skills<br>(T) Working with and formatting<br>(T) Repository skills<br>(T) Access, sharing, dissemination skills |
| Ownership arbitrator | • Shaping the data policy.<br>• Developing access procedures and processes.<br>• Develop documentation that guide the institution in the | (T) Administrative skills<br>(T) Developing policy & procedural documents<br>(T) Appraisal skills<br>(T) Licensing of data skills<br>(T) Citation of data skills<br>(T) Data usage skills |

| **Roles of the data steward (refer to 3.2)** | **Responsibilities (refer to 3.2)** | **Skills necessary to address the responsibility identified (refer to 3.3 – 3.6)** |
|---|---|---|
| | creation, collection and consumption of the data.<br>• Licensing data.<br>• Guiding the proper citation of data. | |
| Data migrator | • Making modifications to the data infrastructure (which includes technology migration for new data representation).<br>• Migration verification (which includes ensuring that queries can be re-executed correctly for the verification of successful data and query migration). | (T) Administrative skills<br>(T) Developing policy & procedural documents<br>(T) Project management skills<br>(T) Data usage skills<br>(T) Practical and technical data skills |
| Preservation manager | Archiving data.<br>Managing data storage. | (T) Administrative skills<br>(T) Developing policy & procedural documents<br>(T) Appraisal skills<br>(T) Archiving skills<br>(T) Metadata skills<br>(T) Project management skills<br>(T) Working with and formatting<br>(T) Managing data storage<br>(T) Repository skills<br>(T) Organisation skills<br>(T) Practical and technical data skills |
| Project manager | • Project planning and management communicating new and changed business requirements to individuals affected.<br>• Managing time and deadlines.<br>• Working independently.<br>• Developing budgets. | (S) Time management<br>(S) Independent worker<br>(S) Attention to detail<br>(S) Building relationships<br>(S) Developing budgets<br>(T) Project management skills |

**Legend**

| **Skills necessary to address the responsibility identified** |
|---|
| (S) – Soft skill |
| (T) – Technical skill |
| (K) – Knowledge component |

With the identification of big data steward responsibilities, it is important to link these responsibilities to the skills identified in this chapter, to gain a clearer understanding of which skills are needed to fulfil specific responsibilities, and how these responsibilities then fulfil the role of a big data steward.

## 3.9 Conclusion

The literature and illustrations in this chapter, as well as in those Chapter 2, serve as important foundation chapters for the rest of this research study. The theoretical content discussed in Chapters 2 and 3 is important with regard to setting the benchmark for the development of a curriculum for big data stewardship training in South Africa (refer to Chapter 6.3). The next chapter explains the methodology that was used to check whether the international academic community has identified the same competencies to address in their big data stewardship education and training efforts.

# Chapter 4

## 4. Research methodology

### 4.1 Introduction

This chapter contains a description of the methodology used in this study. Other aspects discussed in detail, include the research approach followed in the study, the research site, the target population, and the data collection techniques that were considered and used. The data collection instrument used is defined, as well as the method used for data analysis. The validity and reliability of the methodology, and the limitations of the study, are also provided. Lastly, ethical clearance and conduct are discussed before the chapter is concluded.

### 4.2 Research methodology

This chapter contains information regarding the research methodology selected for this research. In this chapter, the researcher elaborates on the research approach selected, the research method selected, and the research site where the research took place. An understanding as to why the research approach, research method, and research site were selected and were most appropriate for this research is presented.

#### 4.2.1 Research approach

When considering a research approach to follow, it is important to consider all the possible approaches, such as the qualitative approach, the quantitative approach, and the mixed-methods approach. The first approach to be discussed in detail is the qualitative research approach.

##### 4.2.1.1 Qualitative research

Qualitative research can be defined as a holistic and emergent approach, as it includes a specific focus, design, measurement tools, and interpretations which are constantly developing and may alter in the course of the research (Leedy & Ormrod, 2015: 99). In a qualitative research environment, the researcher approaches the situation at hand with an open mind and is willing to interact with participants on a personal level. The different variables in qualitative research emerge from the data, which result in finding relevant

information for the study, and, ultimately, to theories explaining the phenomenon being studied (Leedy & Ormrod, 2015: 99). A qualitative approach is therefore characterised by the aim of the study, the methods, which will generate the necessary information for the study and the in-depth understanding of the data (Brikci & Green, 2007: 2; Leedy & Ormrod, 2015: 99). A researcher following a qualitative research approach aims to understand the experiences and attitudes of the participants in the study and to know their personal characteristics, while the participants may know the researcher and their biases (Leedy & Ormrod, 2015: 100 and motivated by Oflazoglu (2017).

There are four main advantages to using a qualitative research approach, the first is that qualitative research allows for rich and in-depth detail pertaining to the study, much more so than is the case with a quantitative research approach. This is because participants are given the opportunity to elaborate in detail when the research is being conducted, taking the research and the participants of the research to a more personal level and a deeper understanding (Leedy & Ormrod, 2015: 282). The second advantage is that the research can also focus on human factors. This means that the perceptions and emotions of the participants regarding the study at hand may be considered. This can be regarded as bias, depending on whether the researcher transgresses the boundaries of ethical research (Leedy & Ormrod, 2015: 282). The third advantage is that a qualitative research approach is appropriate for situations and studies where a detailed understanding of the situation at hand is required. Qualitative methods usually give the researcher more insight into the situation, the research topic, and the participants (Leedy & Ormrod, 2015: 282). The last advantage is that qualitative research allows the researcher to look at the problem more holistically, considering all the relevant factors pertaining to the situation at hand.It is important to note that the more facts there are, the greater the influence on the outcomes (Leedy & Ormrod, 2015: 282).

One must also consider the disadvantages of qualitative research, in order to gain a clear understanding of the approach. The first disadvantage is that the sample sizes of the research group may be small. This means that generalised assumptions and statements cannot be made regarding the data. It also means that the research is of a subjective nature, although more detailed than a quick quantitative survey (Leedy & Ormrod, 2015: 283). The second disadvantage is that both the conditions and the situation in which the research took place

have a limiting influence on the conclusions reached from the data. (Leedy & Ormrod, 2015: 283) and this fact must be kept in mind by the researcher. The last disadvantage is that because of the personal nature of the research, different results may be gained from different participants, or from the same participants on different days when different influencing factors may exist, these being personal or impersonal factors, perhaps causing the outcome of the research to appear inconclusive or less reliable (Leedy & Ormrod, 2015: 283).

#### 4.2.1.2 Quantitative research

The next research approach to consider is the quantitative approach. Quantitative research focusses on numbers. The participant group is usually large, and the results may be generalised to a larger target group. Statistical patterns often manifest themselves in quantitative research, and from these generally applicable conclusions can be reached (Leedy & Ormrod, 2015: 98; Chu, 2015: 36-41).

Quantitative research involves larger sets of participants than is the case with qualitative research and consequently leads to larger sets of data. Statistical techniques are often used to analyse the data. This approach is focussed more on a macro view than on the micro view of a given situation (Leedy & Ormrod, 2015: 98; Oflazoglu, 2017).

A researcher following this approach often finds explanations and predictions in the data captured from the large group of participants that may be generalised. The researcher's intent would then be to recognise any relationships and commonalities among the larger amount of data, which could confirm the predictions and assumptions generally made regarding the data (Leedy & Ormrod, 2015: 98). In this case, the researcher would then collect the necessary data from the large sample group. The data, which are then collected and analysed efficiently, will represent the target population at hand (Leedy & Ormrod, 2015: p99). Furthermore, a researcher following the quantitative approach reduces the data gathered from the large number of participants and summarises them before analysing and making more sense of them. The researcher often calculates average statistics of the greater statistics of the data gathered, meaning that the results are generally in the form of a report and the data are in the form of statistics and impersonal language (Leedy & Ormrod, 2015: 100).

Leedy and Ormrod (2015: 100) note three advantages of quantitative research, the first advantage is that quantitative research consists of a large sample size, which allows the researcher to generalise the conclusions reached. Secondly, that the statistical analysis associated with quantitative research is seen as reliable and, thirdly, that quantitative research allows for systematic and standardised comparisons.

There are three disadvantages of quantitative research according to Leedy and Ormrod (2015: 101), firstly, the fact that quantitative research excludes the human experience and personal perceptions as it focusses more on the environment and situation rather than on personal attributes. The second disadvantage is that quantitative research is limited in that it cannot provide answers to questions such as “Why?” or “How?” regarding the research, and, thirdly, that it focusses on the majority of the target population, which means that the minority are disregarded because their statistics are not large enough to be generalised and considered.

#### 4.2.1.3 Mixed-methods research

The last research approach to be considered is the mixed-methods approach. The mixed-methods approach can be defined as a mixture of qualitative and quantitative qualities to form one approach and one outcome with regard to the research. Furthermore, it can be defined as an approach that combines the elements of the two approaches for the purpose of gaining more breadth and depth in both the research and the results of the data, thereby enhancing the understanding and corroboration of the data collected for analysis (Schoonenboom & Johnson, 2017: 108). Importantly, the mixed-methods approach is used to conduct research on any given field relevant to the research at hand. The researcher gathers both quantitative and qualitative data and integrates both the two methods, and the data gathered. The researcher uses the combined strengths of both sets of gathered data, to understand the research problem fully (Creswell, 2015: 2). Accordingly, the researcher combines quantitative and qualitative data to strengthen the results of the research, to enhance the depth and insight of the research and to answer the research question more fully, instead of drawing conclusions from the data derived from a single approach only (Creswell, 2015: 2).

Creswell (2015: 6) identified three methods with regard to mixed methods, namely, the convergent design, the explanatory sequential design, as well as the exploratory sequential

design. The convergent design can be defined as collecting both quantitative and qualitative data for analysis, and then the researcher must merge the results from the analysis for the purpose of comparing the results with one another (Creswell, 2015: 36). The use of a convergent design can be advantageous for a researcher who would like to gather both quantitative and qualitative data during their data collection process, as this ensures that both forms of data are brought together and it gives the researcher an opportunity to view the data from many different angles (Creswell, 2015: 37).

The second methods associated with mixed methods identified by Creswell (2015: 37), is the explanatory sequential design, which can be defined as the researcher beginning the data collection process by only using a quantitative method. The quantitative method will be used for data collection as well as data analysis. The qualitative method will then be used at the end of the process to explain the results of the quantitative method. The explanatory sequential design is advantageous as both the quantitative and qualitative methods work together and build upon one another to create distinct and easily recognisable stages of the process (Creswell, 2015: 38).

The third and final method identified by Creswell (2015: 39) used in mixed- methods is of the exploratory sequential design. The exploratory sequential design consists of three phases through which the researcher must go; the first phase involves qualitative data collection and analysis, the second phase involves taking the qualitative results from the first phase and transforming them into a new form or instrument appropriate for an experiment with involves quantitative methods. The third and final phase involves quantitative methods in order to apply the new form or instrument to test it or using the progress in the first and second phase for a new experiment. The exploratory sequential design is advantageous as the first phase, being qualitative, submits the raw results in whichever form they come. These results can then be transitioned by means of quantitative methods to make more sense of the results; with this, the results can also be transformed accordingly to suit the need or context of the result. This design is versatile and can change accordingly in form (Creswell, 2015: 40).

The advantages and disadvantages of the mixed-methods approach are the advantages and disadvantages of the qualitative and quantitative approaches. The advantages and

disadvantages already discussed are applicable to the mixed-methods approach, these, of course, varying with regard to relevance from situation to situation.

Having considered all three the research approaches mentioned above, the approach applied throughout this research, and which has been deemed to be most applicable and most advantageous for this study, is the qualitative research approach.

### 4.2.2 Research method

The research method used in this study is content analysis. “A content analysis is a detailed and systematic examination of the contents of a particular body of material for the purpose of identifying patterns, themes, or biases within that material” (Leedy & Ormrod, 2015: 102). A content analysis requires a great deal of planning and preparation before the researcher can begin the analysis of the data (Leedy & Ormrod, 2015: 275 and Chu, 2015: 36-41). In order to understand the nature of a content analysis fully, as well as how it is applied, the steps taken by the researcher are presented below.

During the first step of the content analysis process, the researcher has to identify the material and data to be analysed clearly. Depending on the size of the material and data, it can either be studied in its entirety or by means of selecting a sample (Leedy & Ormrod, 2015: 275). During the next step, the researcher defines the characteristics of the study to make it clear what specifically must be taken note of and done. During this step, the researcher may make use of examples of what the characteristics may look like in the study, to make the definition of these characteristics clearer (Leedy & Ormrod, 2015: 275). During the third step, the researcher breaks the data and material down into smaller segments to make it easier to analyse, if the material is too large or too complex to be understood and managed simultaneously (Leedy & Ormrod, 2015: 275). During the last step of the content analysis research process is for the researcher to analyse the material and data critically. The researcher needs to make sense of what is seen in the data. This sense-making happens at different levels of complexity and in varying segments depending on the size of the material or data (Leedy & Ormrod, 2015: 275).

Content analysis is most applicable for this, as it involves the analysis of existing data (web site content), the purpose is to make judgments and recommendations based upon the given data. A “comparison like” technique is almost always applied during content analysis. For this

study, existing big data curricula, used at different tertiary institutions, locally and abroad, are compared to create the recommended structure for the UP. The data collection technique involves searching the Internet and documenting the data using a data sheet (see Appendix 1). Once all the data were collected, segments were transferred to data tables as this was seen as the most appropriate method to capture the data efficiently. It also made the comparison of the data clearer, thus making judgment calls and the analysis of the data easier for the researcher.

### 4.2.3 Research site

This research, even though it includes data and information from tertiary institutions both locally and internationally, was conducted at the University of Pretoria. This was an exploratory exercise, and it was not necessary for the researcher to travel to all the identified tertiary institutions. Instead, the researcher collected the data online. E-mail communication with other chosen tertiary institutions was not planned and was also not executed. It was anticipated that online communication may perhaps be necessary with the relevant tertiary institutions to gain more clarity on the data collected. However, the specific data were readily available and easily accessible without needing communication with the institutions. Data were therefore only collected from official websites (refer to Appendix B).

## 4.3 Target population and sampling

A target population is defined as the total group of individuals or participants from which the sampling of the data is gathered. The target population is selected because of its relevance to the research. Sampling techniques are then applied to the selected target population to identify the candidates to contact (McLeod, 2014). A target population is thus the entire set of elements and parts that could be used to make inferences, analyse and reach conclusions regarding the study at hand. The target population is specifically defined and selected so that the findings of the research can be generalised and made applicable to the selected group (McLeod, 2014).

Sampling is defined as the process of selecting a particular and applicable subset of the target population of the study (McLeod, 2014). Different sampling methods exist. These are discussed in order to make a clear and credible selection of at least one of the methods for

the study at hand. Sampling methods include random sampling, stratified sampling, volunteer sampling, and opportunity sampling.

Random sampling is defined as the case in which every participant within the selected target group has an equal chance of being selected for the study. This method usually ensures an unbiased selection of candidates (McLeod, 2014). The advantage of using random sampling is that bias is eliminated in the study, this is the case, if the selection of the target population is of an unbiased nature. The disadvantage is that, depending on the size of the target population, and depending on the researcher's resources for the study, an important aspect can be missed because the candidates who have the answers were not chosen randomly (McLeod, 2014).

The second type of sampling technique is stratified sampling. In stratified sampling, the researcher divides the target population into segments so that the researcher can identify the different proportions needed for the sample to be of fair representation (McLeod, 2014). An advantage of stratified sampling is that the outcome of the sampling method is more accurate than pure random sampling in terms of the data being a true representation of the target population. A disadvantage is that stratified sampling can be extremely time consuming and difficult to do well if the researcher does not know the population characteristics in much detail (McLeod, 2014).

The third type of sampling is opportunity sampling. Opportunity sampling is defined as a convenience type of sampling, as it only includes participants from the target population who are, firstly, available to take part in the research and, secondly, are willing to take part (McLeod, 2014). An advantage of opportunity sampling is that it is a quick and easy method of selecting participants, but a disadvantage is that the research could be biased, and the outcome may not represent the target population accurately, thus jeopardising the research study (McLeod, 2014).

The last type of sampling method is systematic sampling. In systematic sampling, the researcher chooses participants within the target population in an orderly and logical way and for a particular reason (McLeod, 2014). The researcher has a free choice of the participants of the study and chooses each participant for a specific reason. Systematic sampling is advantageous, as it should be a true representative sample from the target population that is

valid and applicable to the study. A disadvantage is that it can be time-consuming (McLeod, 2014).

The sampling technique applied in this research is systematic sampling. The participants within the target population were chosen systematically for the value they were expected to contribute to the outcome of the study.

The selected target population for this research comprises certain tertiary institutions in South Africa and selected universities from universities abroad. The researcher wanted to see what the training programmes for big data curators and stewards contained. The tertiary institutions were selected because there is big data activity at the institution.

The universities abroad were selected as follows: the top 100 universities on the Times Higher Education list were subdivided by country segment to gain a representative sample from Europe, the United Kingdom, the United States, Australia and South Africa The target population was then expanded to make sure that the top five Library and Information Science and Library and Information Management schools were included.

Those without English websites were eliminated. The top ten in each of the segments were then identified to give a sample population of 33 universities.

The universities sampled from the wider South Africa population are the University of Pretoria, the University of Cape Town, the University of Witwatersrand, the University of the Western Cape and the Sol Plaatje University. These universities were selected because of their academic status in terms of the research topic, in the end, only the University of Pretoria, the University of Witwatersrand, and the University of Cape Town were used. The other two universities were discarded because of the lack of content and data on their websites.

### 4.4 Data collection techniques

In this case, data collection involves desk research. The internet, websites, and the relevant internet sources were the primary sources of information. These were reviewed for information regarding tertiary training programmes in big data management. Where the information was not uploaded, it was requested via e-mail. The curricula of the big data programmes were specifically of interest. Where available, the skills developed through the programme and any prescribed resources were also identified. It was expected that it would

be possible to use the information collected to create an international "standard" and then make sure that the proposed curriculum is aligned with the standard version.

### 4.5 Data collection instrument

As this research made use of content analysis, data were collected from different internet sources (the websites of specific tertiary institutions). The data were collected using a standardised collection template (see Appendix 1) that indicates the details of the data to be collected. The instrument includes the field in which the data exist, and different parameters of the data, such as the content of the data, the skills being addressed, and any additional data that are required for comparison. The essence of the instrument is captured in the following table.

*Table 4.1: Data collection guide*

| **Template category** | **Motivation for inclusion** |
|---|---|
| Field of study (Natural Science / Economics / Humanities) | |
| Modules of data-related study included in curriculum (Year of study / credits) | |
| Content of data management curriculum | |
| Prescribed work | |
| Skills development activities | |
| Learning outcomes | |
| Content of big data curriculum | |
| Prescribed work | |
| Skill development activities | |
| Learning outcomes | |

| Template category | Motivation for inclusion |
|---|---|
| Additional notes | |

### 4.6 Data analysis

The data analysis of this research is content analysis that is defined and discussed in detail under heading 4.2.2. In this section, the advantages and disadvantages of content analysis are discussed, to show why it is the most useful choice for this study.

The advantages of content analysis are that it allows for both qualitative and quantitative reporting. This means that whichever one suits the study best can be applied, and the research is not restricted to one type of analysis (Leedy & Ormrod, 2015:201). Content analysis is also advantageous as it provides more contextual data in terms of insights throughout the research process, which assists the researcher to gain better insight into the data (Leedy & Ormrod, 2015: 201). Content analysis also allows alternation between different categories and relationships of the data, depending on the context in which the data exists (Leedy & Ormrod, 2015: 201). Another advantage of content analysis is that it is unobtrusive with regard to the interactions with the data (Leedy & Ormrod, 2015: 201; Chu, 2015: 36-41).

The disadvantages of content analysis include that it can be more time-consuming for the researcher than other data analysis techniques (Leedy & Ormrod, 2015:202). Content analysis can also be disadvantageous if a higher level of interpretation is needed (Leedy & Ormrod, 2015: 202). Another disadvantage is that content analysis often lacks a theoretical base or framework, as it is more practical in terms of drawing meaningful inferences regarding the relationships and impacts which the researcher creates (Leedy & Ormrod, 2015: 202). Furthermore, content analysis can be difficult to automate and capture the data accurately, and at times may disregard the context in which the data are produced and focus only on the data itself (Leedy & Ormrod, 2015: 202).

With all the above advantages and disadvantages regarding content analysis, it was still the most applicable data analysis technique to use for this research.

## 4.7 Validity and reliability

This research will be difficult to repeat and gain the same results as the research is based on information valid at a specific time. The data may change, due to possible updates or content removal. The interpretation of the research data may vary from researcher to researcher. Future researchers could rather take the given recommendations into consideration when conducting research, as developments in the field may have taken place that influence the validity and reliability of the study.

## 4.8 Limitations of the methodology

Four limitations were identified. The first limitation of the study is that the entire methodology hinges on the availability of information publicly on the Internet. Although the top 100 universities were identified, there is no guarantee that all these universities conduct big data training. Similarly, there is no guarantee that all universities provide their curricula online. For the South African universities only, the research-intensive universities and Sol Plaatje University were selected. There is no guarantee that their big data activities are provided online.

Another limitation is that the researcher did not have the convenience of time so that the researcher could contact a knowledgeable person at each of the target institutions. The research was, therefore, limited in terms of how much data could be retrieved via the internet and through online communication with the given contact person. The researcher may thus not have been able to gain access to all the available and relevant data.

The third limitation was a language barrier. The research was conducted in English and only websites with English content, were consulted. Those institutions that may have relevant training but do not disclose the facts in English were not accessed.

The fourth limitation is that, because the field of study is so dynamic, it was acknowledged that all the literature and website content published after the dates, as listed in Appendix 1, were not consulted. Accordingly, the researcher could not monitor all the sources published after the given dates.

Other than the above-mentioned limitations, there should be no other limitations in this study.

## 4.9 Ethical considerations

The information considered and the data collected are all in the public domain. Hence, it was not necessary to request ethics clearance. The researcher applied due diligence to ensure that the information collected was treated ethically, so that there would not be misrepresentation. A list of the websites consulted is attached (see Appendix 1).

## 4.10 Conclusion

This chapter discussed the research methodology that was applied to ensure the successful collection of relevant research data. This chapter assists in bringing the research and content of the entire study together and in creating a successful and finalised product of research.

# Chapter 5

## 5. Findings and analysis

### 5.1 Introduction

In this chapter, the data collection instrument (see Appendix 1) is discussed and explained to illustrate how and where the data were retrieved. The chapter then reports on the collected data in detail showing the differences and similarities in the way that the different universities and the various Library and Information Science schools, both abroad and in South Africa, treat the teaching of big data curation. The analysis of the data then led to recommendations regarding the route the UP Department of Information Science could follow (these recommendations are explained fully in Chapter 6). This chapter, therefore, acts as a tool and reference for Chapter 6, where appropriate recommendations are then made for future academic research and further use of the data.

Furthermore, the purpose of this chapter is to address two of the research questions (see section 1.3), namely:

- How are a selected group of international university departments and Library and Information Science schools approaching the education of big data stewards?
- Judging from online content - how is big data stewardship education being addressed by South African university departments and Library and Information Science schools?

### 5.2 Findings

This section will report on the data which were gathered from each of the university departments and LIS schools. The data were illustrated in the best-suited way to aid with understanding and interpreting the data successfully.

#### 5.2.1 Institutions consulted

A conscious decision was made to select a representative sample from the following leaders in data management: Australia, Europe, United Kingdom, as well as the United States and after that leaders from South Africa in data management were also included in the study. Five universities from each region were selected based on their international ranking (as was

explained in section 4.3) and then whether they had relevant information regarding big data curation training on their websites. The sample was then expanded to make sure that the top five Library and Information Science and Library and Information Management schools were included. The term 'institutions' was used to indicate both universities and LIS / LIM schools. Where the data were analysed separately, it is clearly indicated that this is the case.

With the above stated, it is also important to note that many more university websites were analysed and then rejected because there were no data to be collected, regarding the topic, from the site and therefore the institution was not selected for the study. Another reason why some institutions were not used was because of the language barrier - where it would simply have been too costly to have translated the data properly. The institutions which were selected, and from which data were collected, did indeed hold valuable data for the study. The institutions on which are reported, are the institutions which currently have relevant information on their websites.

The selected institution, the web address from where the data were collected, as well as when the collected data are reflected in Appendix 2. The date when the data were accessed and collected is important, as it is acknowledged that the content may have changed/been updated after the collection date. All-in-all, 42 websites were consulted from which 35 were selected – five from the United States, five from Australia, five from Europe, five from the United Kingdom, five from South Africa, and the ten Library and Information institutions. It is important to note that the United Kingdom data were separated from the European data, simply because the United Kingdom alone had enough Library and Information institutions with rich data to make up their own section for analysis. The reason why the remaining seven universities that were then not used for the study were either because of the language barrier of the website of the university, or simply because the data needed for the study were inaccessible or did not exist on the website. This then gave the reason to rule out the use of the specific university.

Furthermore, after the above was determined in terms of which universities would be used for the study and which universities would be taken out of the target group, it is important to note that only three of the five selected universities from South Africa were used in the end. The reason for this is because only three of the five South African universities provided data

regarding the topic of the study, the remaining two did not contain any valid data for the study. In the end, the total number of institutions then used was 33.

Another important factor of which to take note, is that all the top ten Library and Information Science Schools are based in the United States of America. This has caused a predictable but unintended bias, which will have to be rectified in a further study.

### 5.2.2 Field of study

In terms of the field of study from which the data were gathered, this will include from which faculty or department the data were gathered and in which field of study the data exist within the specific institution. While gathering the data it was established that there is no uniformity when it comes to the discipline that hosts the relevant data management degree. From the data gathered across all 33 institutions, the disciplines were diverse in terms of: information management, data science, big data science/management, computer science, information studies, information and library science, engineering, marketing, information technology, statistics, medical sciences, politics, as well as informatics.

With regard to the universities, the data were usually gathered from the different fields of Information Technology studies, namely information management, data science, big data management, computer science, information and data management, as well as information studies. There were exceptions where the information was extracted from fields as diverse as Economics, ICT, Medicine and Engineering.

With regard to the Library and Information Science and Library and Information Management schools, the data were gathered from a variety of Departments namely computer science, information and library science, information science, information management, data science, information studies, as well as library, archival and information studies. The table below summarises the diversity.

*Table 5.1: Regional analysis of departmental distribution*

| Region | Departments where data were extracted |
|---|---|
| Australia | • Business, Economics and Statistics (big data)<br>• Data Science (big data)<br>• Big Data (big data)<br>• Engineering, Architecture and Information Technology (long tail data)<br>• Data Science (long tail data) |
| Europe | • Big Data (big data)<br>• Informatics (long tail data)<br>• Computer Science (long tail data)<br>• Business and Economics (big data)<br>• Big data analysis (big data) |
| South African | • Big Data Science (long tail data)<br>• Data Science (long tail data)<br>• Big Data Analytics (long tail data) |
| United Kingdom | • Engineering (big data)<br>• Medicine (big data)<br>• Political Science (big data)<br>• Data Science (long tail data)<br>• Information Technology (big data) |
| United States | • Data, systems and society (big data)<br>• Big data (big data)<br>• Engineering (big data)<br>• Marketing (big data)<br>• Information Science (long tail data) |

Note: The table above also states in brackets whether this comes from a big data or long tail data curriculum.

From the above, it is clear that the most common field of study from which the data were gathered and where the data existed was that of big data science/management, the second most common field of study was that of data science, and the third most common field of study was computer science and information and library science.

Figure 5.1 below illustrates all the fields of study interested in big data curation.

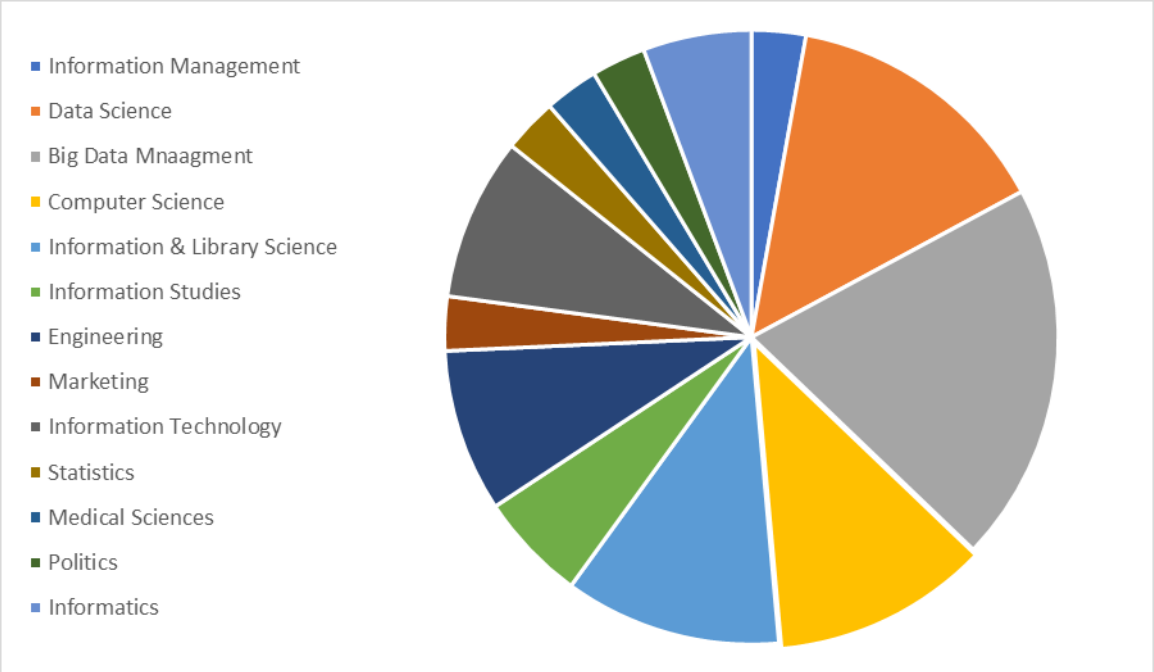


Figure 5.1: Field of study where the curriculum was represented

It is important to note that throughout this part of the data collection process, it was easier to trace more data specific courses than those of big data courses. It took in-depth searching to locate the big data curricula.

The chart can be read from the list on the left of the chart from top to bottom, beginning with 'Information Management' which is indicted in the dark blue colour on the top (slightly right) of the chart, through to 'Informatics' which is on the top (slightly left) of the chart.

### 5.2.3 Module of data-related study (academic level)

The module of data-related study includes at which level of degree the data existed, this would include whether or not the data pertained to an undergraduate level, or a postgraduate

level. The module of the data-related study included all levels of study, from a certificate of diploma to doctoral level. Most of the data which were collected with regard to this section were at a master's level of study, this being followed by a post-graduate level of study where only an undergraduate degree was needed as a prerequisite for the specific course. The third most prevalent module of the data-related study was that of a certificate or diploma, this being followed by an honours level of study as well as a doctorate level of study. The statistics of this data are illustrated in Figure 5.10 below:

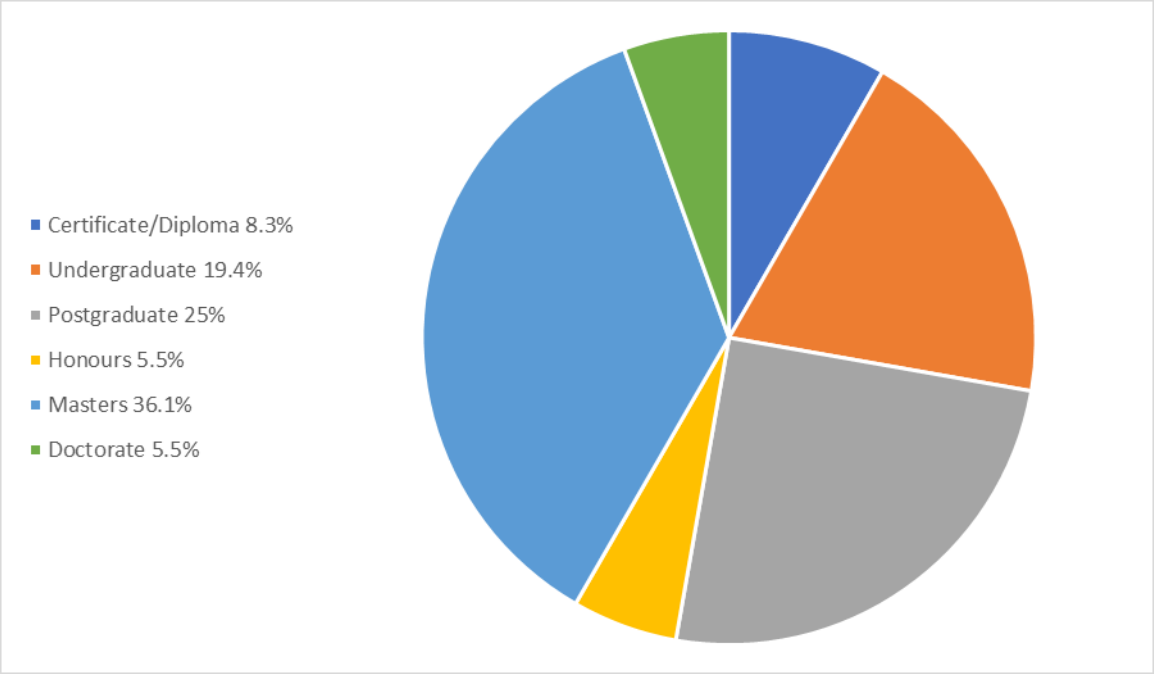


Figure 5.2: Module of data-related study

The chart above can be read in the same fashion as Figure 5.1, where the labels on the left are read clockwise from the top following the colour-legend. It is important to note that the majority of the courses which were identified for the purpose of this study, contained the relevant curricula at master's level, with the second highest number of courses being at postgraduate level, and the third highest statistic was at undergraduate level. It was assumed that the focus is at master's level as that is when researchers get to understand the intricacy of the curriculum content completely. This is also when students need to demonstrate more advanced research capabilities. Very often they have, by then, acquired learning and research styles that enable them to appreciate the value of the data or big data management module or course.

### 5.2.4 Content of data management curriculum

The content of the data management curriculum includes what was found in the data collection process in terms of the modules they provide and the number of credits each module carries. This category will also include if the modules are core-modules, or if they are electives. More details will be provided on this section in Appendix 3 and Appendix 4 where all the data will be tabulated for a better understanding of the content.

The Library and Information Science Schools provided modules on a variety of levels, from core-modules at a master's degree level, to modules at an undergraduate level. Modules varied from data management strategies, data science principles, information analytics, database management and an overview thereof, as well as data modelling, data analytics as well as an introduction to basic programming. Accordingly, most of the curricula included data management and general database management, all the other modules noted in these curricula included data management.

The Library and Information Science schools also provided core-modules as well as elective based modules. The majority of the curricula which were found were at postgraduate level. The Library and Information Management schools produced modules, which were database related, data analytic related, as well as more technical modules, such as machine learning, computer security, data modelling, and introduction to programming languages. It is important to note that these institutions provided a more technical overview of modules including more hands-on learning and software learning associated with the above-mentioned modules.

The institutions from the United States had more curricula regarding big data than those on data management, but nonetheless two of the identified institutions provided curricula based on data management. The identified modules were at a master's level, as well as that of a certificate. Once again, database learning was included in these modules, along with actionable data, data analytics, as well as ethics in data management. As mentioned previously, more curricula were found on big data management than on data management, but the data management modules did include the important modules of database management and data related concepts as mentioned above.

The institutions from Australia, which were included in this study, regarding data management curricula provided courses at a master's level and a few at undergraduate level. With regard to data management curricula, only two institutions could provide curricula, whereas, the other institutions provided curricula on big data management. The master's course, in this case, included modules such as an introduction to data science and software engineering, data mining and analytics, as well as electives varying from bioinformatics to algorithms and data structures. The undergraduate course, in this case, included modules based on databases, data science, and data analysis. One can note that between the two institutions, which provided data on data management curricula, the master's level modules are more complicated, and the undergraduate module is more general, based on databases and the data themselves.

The South African institutions all provided data management courses on an honours and master's level. The modules included topics such as database design, data visualisation, data analytics, and machine learning. Electives for these courses varied from algorithms to special topics in computer science which included big data.

From the institutions based in the United Kingdom, only one school from the University of Edinburgh provided a curriculum based on data management, this course was based on a postgraduate course level. The course included modules such as data science, data analytic techniques, as well as data science tools and techniques. The rest of the institutions all provided courses based on big data curricula.

There were only two institutions in Europe that included data management courses, these courses ranged from the graduate level to master's level. The other institutions from Europe provided courses on big data management. The master's data management courses included modules such as data engineering, data analysis, database management, as well as the programming of databases. The graduate degree course included data management courses such as computer science, hardware fundamentals, data storage, networking, as well as systems engineering. It is important to note that the master's level course included more detailed and data specific content, whereas, the graduate level course provided more general based topics.

### 5.2.5 Electives of data management curriculum

No stated and visible electives were found within the data management curriculum of the Library and Information Science schools. All the modules found were indeed a part of the core curriculum.

No electives were found in the data management curriculum other than that of external courses in statistics, data analysis, or big data at the University of British Columbia. Library and Information Science Schools / Departments.

Similarly, no electives were found in the data management curriculum in the United States institutions other than that of the California Institute of Technology (Caltech). The electives included modules such as computational methods, market dynamics, applied remote sensing, project management, engineering management methods, communicating leadership, social behaviour, policy and design, ergonomics and anthropometrics, human-robot interaction, simulation modelling and analysis, leadership theory and practice, as well as management consulting essentials.

The electives which were identified in the Australian institutions came from the data management curricula at the University of Queensland, and the Monash University. The University of Queensland's data management curriculum included a variety of electives namely accounting, bioinformatics, informatics, algorithms and data structures, artificial intelligence, computational science, high-performance computing, software engineering, econometrics, finance, financial mathematics, marketing, consumer research and behaviour, epidemiology, mathematical statistics, probability models, statistical analysis and genetic data, as well as longitudinal and correlated data. Monash University's electives for their data management curriculum included Information Technology for business, as well as software development.

All three the identified South African institutions included electives in their data management curriculum. The University of Pretoria's data management curriculum included electives for their programme such as computer science, informatics, information science, statistics, mathematics, as well as big data science. The electives which exist in the University of Cape Town's data management curriculum include data science for astronomy, data science for particle physics, bioinformatics, data science for industry, decision modelling for prescriptive

analytics, Bayesian decision modelling, as well as data analysis for high frequency trading. Lastly, the electives for the data management curriculum at the University of Witwatersrand included applications of algorithms, computer vision, distributed computing, computing and scientific data management, software defined networking, as well as special topics in computer science such as big data.

There were no identified electives within the data management curriculum at the United Kingdom institutions. More of the electives existed within the big data curriculum at these institutions.

Only one of the institutions included electives in their data management curriculum in the European institutions. The Technical University of Munich includes electives such as data engineering, advanced topics in data engineering, data analytics, and data analysis as electives for their data management module.

### 5.2.6 Skills being addressed in data management curriculum

With regard to the skills being addressed in the data management curricula of the selected target group, it is important to recognise the different courses that were identified, and then adapt the coursework into different skill groups and categories. In order to do this, both responsibilities and skills were identified in Chapter 3 (see sections 3.4 to 3.6). These were mapped with the skills identified within the curricula of the institutions selected for analysis.

To recapitulate: the responsibilities of the data steward were identified, in the literature as follows:

*Table 5.2: Identified responsibilities*

| **Responsibilities of the data steward (see section 3.2)** |
|---|
| • Facilitating and managing the appropriate and efficient use of the big data assets.<br>• Ensuring that data is FAIR (findable, accessible, interoperable and reusable). |
| • Working collaboratively in teams.<br>• Building relationships with other data stewards.<br>• Collaborating with other data stewards to establish data standards where these may not exist. |

| **Responsibilities of the data steward (see section 3.2)** |
| --- |
| • Interpreting heterogeneity in big data.<br>• Managing inconsistency and incompleteness in the data set.<br>• Being able to address issues of scale in the big data set.<br>• Making use of real-time techniques to keep up to date with the ever-growing amounts of data so that data overload does not occur.<br>• Managing the privacy and ownership of big data without compromising the value of the data.<br>• Finding ways to visualise data so that the data is useful from the human perspective.<br>• Organising data (skills to illustrate understanding of the data and the purpose of the data). |
| Maintaining expert level knowledge about:<br>• Big data life cycles.<br>• Funder requirements.<br>• The value of big data.<br>• Discipline knowledge, including discipline-specific methodologies.<br>• Safety, security licensing and copyright of data.<br>• Developing a good understanding of research ethics.<br>• Understand the importance of data as a secondary resource. |
| • Facilitation, data presentation, understanding of learning styles, and working with different types of people<br>• Training colleagues and researchers on the following topics:<br>  ○ Naming conventions.<br>  ○ Versioning.<br>  ○ Labelling.<br>  ○ Writing data management plans.<br>  ○ Writing data documentation.<br>  ○ Allocating metadata. |

| **Responsibilities of the data steward (see section 3.2)** |
|---|
| • Developing and promoting data quality standards<br>• Ensuring that the data are associated with the correct metadata<br>• Ensuring availability of the data and ensuring accessibility to the data<br>• Considering the timeliness of the data<br>• Checking the authorisation of the data<br>• Ensuring **usability** (which means documentation of the data, credibility of the data, and the data's metadata)<br>• Taking responsibility for the **reliability** of (trust in) of the data (which includes accuracy of the data, integrity of the data, consistency mechanisms of the data, completeness of the data, and auditability of the data)<br>• Taking care of **relevance** issues (which includes the degree of correlation between data and content, and the data users' expectations of the given data, being able to advise on the adaptability of the data to the specific need of the user<br>• Providing a valid description method for the given data, (which then allows for a full understanding of the data by the users) |
| • Facilitating and managing the appropriate and efficient use of the big data<br>• Managing access to data<br>• Establishing and maintaining repositories<br>• Controlling data security<br>• Ensuring the persistent identification of data sets<br>• Measuring the impact of data sets |

| **Responsibilities of the data steward (see section 3.2)** |
|---|
| • Providing machine actionable landing pages enable the accessing of metadata and data via query re-execution<br>• Considering query uniqueness<br>• Ensuring that the sorting of data sets is unambiguous and reproducible<br>• Setting up a standardised result set verification system<br>• Timestamping queries (when they are resolved compared to when they were made)<br>• Query identification<br>• Storing query metadata<br>• Creating automated citation texts |
| • Shaping the data policy<br>• Developing access procedures and processes<br>• Develop documentation that guide the institution in the creation, collection and consumption of the data<br>• Licensing data<br>• Guiding the proper citation of data |
| • Making modifications to the data infrastructure (which includes technology migration for new data representation)<br>• Migration verification (which includes ensuring that queries can be re-executed correctly for the verification of successful data and query migration) |
| • Archiving data<br>• Managing data storage |
| • Project planning and management<br>• Communicating new and changed business requirements to individuals affected<br>• Managing time and deadlines<br>• Working independently<br>• Developing budgets |

The identified skills were the following:

*Table 5.3: Identified skills*

| Knowledge components (Refer to section 3.4) | Technical skills (Refer to section 3.5) | Soft skills (Refer to section 3.6) |
|---|---|---|
| Data life cycles (3.4.1) | Writing DMPs (3.5.1) | Time management (3.6.1) |
| Funder requirements (3.4.2) | Administrative skills (3.5.2) | Independent worker (3.6.2) |
| Value of data assets (3.4.3) | Developing policy & procedural documents (3.5.3) | Attention to detail (3.6.3) |
| Discipline-specific requirements (3.4.4&5) | Appraisal skills (3.5.4) | Building relationships (3.6.4) |
| Safety and security (3.4.6) | Licensing of data skills (3.5.5) | Community-based data skills (3.6.5) |
| Licensing and copyright (3.4.7) | Archiving skills (3.5.6) | Developing budgets (3.6.6) |
| Research ethics (3.4.8) | Metadata skills (3.5.7) | |
| Secondary data use (3.4.9) | Citation of data skills (3.5.8) | |
| | Project management skills (3.5.9) | |
| | Data usage skills (3.5.10) | |
| | Data research skills (3.5.11) | |
| | Working with data and data formatting (3.5.12) | |
| | Managing data storage (3.5.13) | |
| | Repository skills (3.5.14) | |

| Knowledge components (Refer to section 3.4) | Technical skills (Refer to section 3.5) | Soft skills (Refer to section 3.6) |
|---|---|---|
| | Organisation skills (3.5.15) | |
| | Access, sharing, dissemination skills (3.5.16) | |
| | Training skills (3.5.17) | |
| | Practical and technical data skills (3.5.18) | |

### 5.2.7 Skills development activities of data management curriculum

No skills development activities were found within the data management curricula of the Library and Information Science Schools as well as in the Library and Information Science Schools. If a practical component existed within the module or curriculum, it would be based on the theory of that module and not for external skills-based development. Likewise, no skills development activities were found in the selected institutions of the United States, Australia, South Africa, the United Kingdom, or in Europe.

### 5.2.8 Content of big data curriculum

The content of the big data curriculum part of this study includes the data that were gathered in terms of the selected tertiary institutions and their big data courses, degrees and programs. This data regarding the big data curriculum were more challenging to find in some tertiary institutions, but then easier to find than that of data management curriculum in other tertiary institutions. This section will discuss all the data which were collected regarding the big data curricula of the selected target groups for this study. More details on the content regarding the big data curricula can also be found in table format in Appendices 3 and 5 where the data are easier to understand and more aesthetically pleasing.

In terms of the Library and Information Management schools, the big data curricula, which were found are of a postgraduate and master's level of study. Although not all the institutions had big data curriculums, the institutions which, did have curriculums included big data content such as introducing big data, big data systems such as SQL, Hadoop and Hive, as well as map-reducing. The rest of the big data curriculum at these institutions included curricula which covered content such as data curation, data models, metadata, laws and ethics, standards, data analytics, algorithmic foundations, as well as big data infrastructure. As one can see, the content identified in the big data curriculum is quite intricate in terms of what is being provided, this can also indicate why it is being provided at a postgraduate level.

The Library and Information Science institutions did indeed have content based on big data curricula, but the data found were extremely broad and not as intricate. It is also important to note that the content provided in terms of the curricula exist at a graduate and undergraduate level of study, which could be the reason why the content is not as detailed

as that of our previous data collected on a postgraduate and masters level. The data which existed in the big data curriculum included content such as tools for using large data sets, database usage via SQL, programming languages associated with big data systems, map-reducing, as well as data modelling, databases, and data streaming systems. The content which was gathered is only relatively in-depth but can still give an overview of what the course is about and the learning content it provides.

The selected institutions from the United States all except one, provided data regarding big data curriculum. The course did, however. vary from certification to a postgraduate degree. The content which was discovered in the courses of certification included unstructured data, recommendation systems, networks and modelling, data security, big data initiatives, ethics, big data models and visualising big data. The certification courses also included big data algorithms, programming, machine learning, and analytic functions. The postgraduate course included content in their big data course such as data systems, data systems architecture, mapping, databases, data skipping, ethics, data provenance, as well as data structure synthesis. Once again, one can note that the content, which was found on the certification level is far less intricate and involved as the big data content found and provided at postgraduate level.

Most of the institutions from Australia provided content on a big data curriculum, with only two of the institutions not providing any kind of big data curriculum. The institutions which did indeed provide content on a big data curriculum offered it from an undergraduate to a master's level of study. The undergraduate level provided bug data content in their curriculum, such as the statistical theory, matrix analysis, and the probability theory. The curriculum also included content of big data features such as heterogeneity, noise accumulation, spurious correlation, and incidental endogeneity. The course, furthermore, also provides study topics such as high-dimensional statistical inference, large covariance matrices, large-scale statistical learning, as well as dimension reduction and component analysis. With the above content, it is rather different to that of other undergraduate institutions which have big data curricula. The content provided for this undergraduate programme is rather more intricate and detailed than that of others we have come across. The master's level of study courses provided data on their curricula including statistical modelling, data mining, statistical techniques, cloud computing, web search and text analysis,

and advanced database systems. Other content for master's level courses included data science, data visualisation, machine learning, visual data analytics, as well as data analytics. The master's level course also provided extremely intricate and detailed content for their course including topics and what the course is about. This has become an expected result for master's courses within this study.

The institutions from South Africa only had one school that provided content in a big data course; this course was also presented at master's level. The course included content such as big data science, machine and statistical learning, data platforms, ethics for big data science, mathematical optimisation for big data science, and big data management. One can note that from the data collected here, the content was comprehensive enough and also followed a more theoretical guide than that of a practical guide. Nevertheless, the topics and content were still good enough to produce a good big data curriculum. This course also came with a varying amount of big data electives where it was assumed that one could go into a more technical side of big data if required.

The institutions from the United Kingdom all provided content on big data courses except for the school from the University of Edinburgh. The level of study for the content of the big data courses varied from undergraduate, to honours, through to a postgraduate level. The undergraduate courses included content in their big data curriculum such as big data analysis, statistics, predictive modelling, rule-based learning, as well as kernel methods. The undergraduate courses also covered content such as data science, statistical learning, computational and machine learning, structured databases, and data preparation and processing. The honours level course included content in their big data curriculum, which covered data processing systems, data analysis, data manipulation, data exploration and visualisation, data mining, and data reporting. The postgraduate courses of study included content in their big data curriculum such as data ethics, modelling and inference, and big data ethics. All the courses mentioned above provided a relatively good amount of content for the big data curriculum and all the data collected are intricate enough to make relevant sense of the data. This is the first time within the data collection process where all levels of institutions form a specific target group, which, in the case of the United Kingdom, provide an intricate and in-depth amount of data.

In terms of the institutions from Europe, three of the institutions provided content on a big data curriculum. Of the three institutions, the courses ranged from all levels including undergraduate, postgraduate, master's, and doctoral level. The undergraduate courses included content such as data systems, database systems, graphics, and social networks. The post-graduate course included content such as data cleaning, statistical methods, big data patterns and outliers, data instruments and devices, systems for big data analysis, as well as machine learning algorithms and data curation. The master's level of study included content such as parallel programming models, big data map-reducing, cloud computing, web services and workflow. The master's programmes also included content such as distributed systems, database management, data mining, advanced statistics, as well as practical artificial intelligence. The doctoral course of study included content such as business analytics and big data, as well quantitative market research. From the data collected above, one can note that the higher level of the course, the more intricate, detailed, and advanced the content became.

### 5.2.9 Electives in big data curriculum

The electives within the big data management curricula of the Library and Information Science institutions only existed at the University of Texas-Austin. The electives included computer vison, natural language processing, as well as bioinformatics. The other institutions did not have any visible electives for documentation.

The Library and Information Science schools' institutions did not have any visible or noted electives for their big data management curricula. This was a similar case to these school's data management curricula where only one school had electives for the programme.

With regard to the institutions from the United States, no electives were found with regard to these institutions' big data curricula.

In the Australian institutions, the University of Sydney and the University of Melbourne both included electives for their big data management curriculum. The University of Sydney's electives for their data management curriculum includes spatial data analytics, data visualisation, business intelligence, predictive analytics, visual data analytics, customer analytics, machine learning, marketing research concepts, as well as statistical learning and data mining. The University of Melbourne's electives for their big data management course includes spatial information, spatial databases, spatial analysis, spatial visualisation, analysis

of high-dimensional data, statistical modelling, mathematics of risk, optimisation for industry, statistics, stochastic calculus, advanced probability, random processes, artificial intelligence, computer science, algorithms, genomics, programming, internet technologies, computing systems, web search and text analysis, knowledge management systems, as well as data warehousing.

The South African institutions did not have any visible electives for their big data management curricula, this should also consider the fact that there were no actual big data curricula identified in these South African institutions, thereby making it impossible to have electives for modules which do not exist.

In the United Kingdom institutions, only one of the institutions contained electives for their big data management curriculum, this was from the University of Oxford, where the electives for the postgraduate to doctoral level courses in big data management included all undergraduate modules within the related field and the same school within the school.

The selected institutions form Europe did not include any electives in their big data management curricula, even though big data curricula do indeed exist in these institutions.

### 5.2.10 Skills addressed in the big data curriculum

The skills that are addressed in the big data curricula of the selected institutions for this study will be represented in the same manner as those that were addressed in the data management curricula discussed section 5.2.6 and Appendix 5 of this chapter. The skills will, therefore, also be mapped to those identified in Chapter 3, section 3.5 and 3.6. The skills are informed by the courses which the institutions offer in their big data curriculum.

The skills are therefore identified within the big data curricula of the selected institutions and will be represented in a table format in order to map the identified and addressed skills to those identified in Chapter 3. This table can be found in Appendix 6 of this chapter.

### 5.2.11 Skill development activities of big data curriculum

The skills development activities that were identified in the big data management curricula of the selected institutions, only existed in institutions in the United States, as well as institutions in Australia. Neither the Library and Information Science institutions nor the Library and

Information Science schools included any skills development activities for their big data management modules. The selected institutions from South Africa, the United Kingdom, as well as Europe also did not contain any skills development activities in their big data management curricula.

The skills development activities which were recognised within the United States institutions, came from the Massachusetts Institute of Technology (MIT). Within MIT's curricula, the skills development activities included instructivism, constructivism, social constructivism, as well as connectivism, which are all the different learning approaches that are practised by teachers and students. Instructivism includes the normal graded tests scenario, constructivism includes students learning through case studies, social constructivism which includes the students learning through social interactions and communication, as well as connectivism which includes discussion groups for knowledge sharing platforms.

The University of Melbourne was the school from Australia, which included skills development activities for their big data management course at the school. The skills development activities included practical modules such as science communication, communication for research scientists, science in institutions, and the science and technology internship. All the components mentioned above are indeed, as mentioned, more practically-based for the students to take in conjunction with the degree core modules.

### 5.2.12 Prescribed work

From the data collected from the Library and Information Science institutions, no prescribed work was identified as part of the curriculum from any of the individual institutions in any of their individual courses:

- The data collected from the Library and Information Management institutions did not contain any data regarding prescribed work for any of the courses offered and identified.
- The data collected from the United States institutions did not include any prescribed work for any of the courses offered and identified.
- The data collected from the Australian institutions did not include any prescribed work for any of the courses offered and identified.

- The data collected from the South African institutions did not include any prescribed work for any of the courses offered and identified.
- With regard to the United Kingdom institutions, two of the institutions included prescribed work in their courses. The first school is the University College London which prescribed work for their undergraduate degree in the Department of Political Science. The prescribed work for the data science module within this course includes students being introduced to quantitative methods, either in statistics or econometrics, at any level. The prescribed work for the course and module also includes familiarity with computer programming or database structures. The second school from the United Kingdom, which entails prescribed work is that from the University of Manchester. The course at the University of Manchester is the BSC Honours course in Information Technology Management for Business – Information Systems in Business and Introduction to Big Data and their Manipulation. The prescribed work includes the student going through prescribed sources, these sources are presented below in Table 5.4

*Table **5.4**: The prescribed sources*

| |
|---|
| Bocij, P., Greasley, A., and Hickie, S. 2008. Business information systems: Technology, development and management. 4th ed. Upper Saddle River, New Jersey: Prentice Hall. ISBN: 027371662X, 9780273716624. |
| Gleick, J. 2012. The Information: A history, a theory, a flood. London: Pantheon Books, Fourth Estate. |
| Harvey, G. 2010. Excel 2010: All-in-one for dummies. [ONLINE]. Available at: https://capdtron.files.wordpress.com/2014/03/excel-2010-all-in-one-for-dummies.pdf. [Accessed: 9 September 2019]. |
| Laudon, K.C., & Laudon, J.P. 2015. Management information systems, managing the digital Firm. 15th global ed. New York:Pearson. |
| McFedries, P. 2013. Formulas and functions: Microsoft Excel 2013. [online]. Available at: http://ptgmedia.pearsoncmg.com/images/9780789748676/samplepages/0789748673.pdf. [Accessed: 9 September 2019].<br>Naughton, J. 2012. From Gutenberg to Zuckerberg: What you really need to know about the internet. London:Quercus. |

Pearlson, K. E., Saunders, C. S., & Galletta, D. 2016. Managing and using information systems: A strategic approach. 6th ed. Hoboken, New Jersey:Wiley. ISBN: 9781119244288.

The data collected from the European institutions did not include any prescribed work for any of the courses offered and identified.

### 5.2.13 Learning outcomes

A table has been created in Appendix 3 to demonstrate the data collected in terms of the learning outcomes for the specific courses and modules from the targeted institutions. It is important to note that not all the sources consulted at the targeted institutions contained learning outcomes, hence, all 33 institutions are not represented on the table.

It is also important to note that the table in Appendix 3 includes the learning outcomes that were identified at the specific institutions. Where no module was listed by an institution in the table, the learning outcome was then stated for the entire course and not a specific module. Where a module was listed and no course was listed, the learning outcomes listed are then appropriate to the module and not the entire course.

It is also important to note that the learning outcomes consisted of long tail as well as big data curricula. One can identify the differences between the long tail data and big data learning outcomes by looking at under which course or module the learning outcome is listed.

Furthermore, from all of the above data collected regarding learning outcomes for the specific courses identified for analysis for this study, it is important to note that most of the institutions included learning outcomes in their courses, but some courses still did not. From this analysis, it is clear that if a module or a course has learning outcomes, it indicates what the course is expected to achieve and what the students will gain from it if they take it. It is also important to note that a learning outcome helps students understand what knowledge they will gain if they take the course and if this is the knowledge that they want to gain and perhaps need to gain. For any curriculum, it is important to include learning outcomes for this reason. It is also then easier to map skills to learning outcomes in order to understand and identify which skills the student will or can gain from taking the specific course or module,

and for the purpose of this study, it is important that those skills are identified and mapped accordingly to the identified learning outcomes.

In Appendix 7, the learning outcomes that were highlighted in Appendix 3 were then grouped in the left-hand column in order to represent the curated learning outcomes. These outcomes that have now been organised and grouped accordingly, are mapped to the specific skills identified in Chapter 3. This mapping of skills, which addressed the specific learning outcomes was an assumption made by the researcher to illustrate the link between the skills as well as the learning outcomes.

It is important to note that all the learning outcomes in Appendix 7 were also mapped with the personal (soft) skills. The personal (soft) skills in Table 5.3 include time management, being an independent worker, having the ability to pay attention to detail, being able to build relationships, having community-based data skills, as well as being able to develop appropriate data related budgets.

The knowledge components that are represented in Table 5.3 of this chapter contained consistencies throughout Appendix 7. The three knowledge components that were repeated and applicable to all the learning outcomes are the value of the data assets, the discipline-specific requirements, as well as secondary data use.

### 5.2.14 Additional notes

The additional note part of the data collection tool includes any additional information that is added to the course for the student to take note of. The data that seems to be important and useful for understanding the course better are then added under additional notes, these are often data that add value to the course. The additional notes, however, could also be information that did not have any other place in the data collection instrument but are too important to leave out.

The additional notes from the Library and Information Science institutions, firstly came from the Syracuse University for their master's degree in information management. The additional notes included different aspects that the course would cover, such as the management of technology, the management of solution development, technical knowledge, the environmental context of information management, professional communication skills,

leadership and teamwork development, as well as information literacy, analysis and problem-solving. The University of Texas – Austin added additional notes for their undergraduate computer science course. The additional notes included prerequisites for the course that indicated that the student should take specific modules before enrolling for the course. These modules were an introduction to big data, advanced data mining, and big data programming. Finally, the University of Maryland – College Park included additional notes for their graduate course in big data systems. The additional notes contained the focus of the course, which is set to be on a diverse set of techniques, tools and systems used for data science on large volumes. The additional notes also indicated that the course covers both relational database systems as well as NoSQL systems. The course goals are aimed at providing an overview of data management systems, focussing more importantly on the strengths and limitations of each system. The course also includes cloud computing and data centres.

The additional notes from the Library and Information Management institutions entailed most of the institutions providing additional notes for their courses. The first school was that of the University of Sheffield for their postgraduate course in computer science. The additional notes pertained to what the course covers and of what it consisted. The course, therefore, covered key techniques for analysing and interpreting data, which were covered by two departments, namely, the Department of Computer Science and the Department of Mathematics and statistics. The course also included a research project based on data analytics which would develop the students' skills in research in the appropriate field.

The second school with additional notes from the Library and Information Management institutions is the University of Illinois – Urbana Champaign for their course in data curation. The additional notes included more information about the course and what the student would be studying. The additional notes were, therefore, an introduction to data curation and the management of data throughout data curation. This also comprised gaining knowledge on the data life cycle as well as broad overview of theoretical and practical problems in the field of data curation.

The third school from the Library and Information Management institutions was that of the University of British Columbia for their data services. The additional notes included the pathway designed for the course, these additional notes included the course developing the competencies necessary to provide services related to data, namely, data curation and

stewardship. This also included a secondary focus on data analysis, as well as the summarising of data visualisation for understanding and communication.

The last school from the Library and Information Management institutions was that of the Indiana University of Bloomington for their master's course in information science, specifically in data science. The additional notes provided an introduction to the course which stated that the programme reached out to multi-disciplinary work in computer science, informatics, information science, as well as engineering. It prepared students for the field and career in data science, and also created the foundation for further studies.

Of the data collected from the United States institutions, three of the institutions contained additional notes for their degrees. The first school was Stanford University for their certification in big data, strategic decisions. The additional notes pointed the key benefits of the course out, these key benefits included the student being able to uncover hidden expectations, correlations, patterns, and trends to bring about better decision-making for the situations at hand.

The second school from the United States with additional notes was Harvard University for their post-graduate course in operating and data systems. The additional notes covered what the student needed to be able to succeed in the course, this included gaining familiarity with basic traditional database architecture, gaining familiarity with modern database architecture, as well as gaining familiarity with the basic modern large-scale systems.

The last school from the United States institutions was the California Institute of Technology (Caltech) for their master's course in data science. The additional notes indicated what the degree could include, this then entailed optional focus areas in data science, interactive technology, network design, as well as market design. The additional notes also stated that data science was the optional focus area for the research and data collection purposes.

From the data collected from the Australian institutions, all the institutions provided additional notes in their courses. The first school was the Australian National University for their undergraduate course in big data statistics. The additional notes indicated what the course would offer the student, this then included an introduction to developments in Random Matrix Theory as well as online learning which may address challenges as well as opportunities created by large amounts of data.

The second school was the University of Melbourne for their master's course in data science. The additional notes included the skill sets that the student would gain from taking the course. The generic skills that could be acquired from the course comprised having the ability to demonstrate critical enquiry, analysis and reflection, having a strong sense of intellectual integrity and ethics, gaining in-depth knowledge in their specialist area, as well reaching a high level of writing and research. The generic skills also included being able to think critically and creatively when approaching specific data problems, having a set of flexible and transferrable skills, as well as being able to initiate and implement constructive change in their communities and professions.

The third school was the University of Sydney for their master's course in big data in business. The additional notes included the overview of the course, which was that the course was designed to provide students with training in big data and analytics. This course focussed on disciplines in business analytics, business information systems, and marketing and the institute of transport and logistics which dealt with big data as well as analytical tools and technologies.

The fourth school from Australia was the University of Queensland for their master's course in data science. The additional notes gave an overview of the course that stated that the course combined studying advances topics from computing, statistics and mathematics, as well as a selection of electives combined with the course core modules. The course also made use of relevant big data technologies and tools as well as developing essential knowledge for ethical use of these tools and technologies.

The last school from Australia was the Monash University for their undergraduate course in data science. The additional notes provided the prerequisites for the course that were modules that worked with Java and Python.

From the data collected form the South African institutions, two of the institutions included additional notes in their courses and data collection. The first school was the University of Cape Town for their master's course in data science. Their additional notes stated that students would learn statistical and computing skills that were required for dealing with big data for astronomy, physics, medicine as well as commerce. The second school was from the University of Witwatersrand for their honours course in big data analytics. The additional

notes included an overview of the course, this entailed that the course would introduce students to the field of big data analytics, because the course had a strong focus on computational and mathematical foundations. Students would also be required to complete a research project for which mentorship would be necessary for the research project, thereby, equipping the students with the appropriate skills needed to conduct research of a high standard.

From the data collected from the United Kingdom institutions, two of the institutions supplied additional notes for their courses. The additional notes included an overview of the module that entailed that tutorials would be provided throughout the module, the module also aimed at providing students with the tools and techniques to understand datasets, as well as how to use these tools to perform data analysis for further data mining.

The second school was the University of Manchester for their honours course in information systems in business and an introduction to big data and their manipulation. The additional notes presented the aims of the course, which was to demonstrate the importance of data within information systems, to demonstrate the data life cycle, as well as to introduce data manipulation, visualisation and analysis. The notes also contained the introduction to the concept of 'big data' and its difference from conventional data, as well as to introduce advanced data processing and analytics techniques through open source languages.

The data collected from the European institutions, four of the five institutions included additional notes for their courses. The first school was the Technical University of Munich for their master's course in data engineering and analytics. The additional notes pointed out what the students would be doing throughout the course, this included handling and analysing large amounts of data which led the students towards the trend of big data and how it was handled. The students would also cover technical advances in examining the large sets of data, as well as the creation of vehicle data and the sharing of information through intelligent networking or intelligent energy grids. The programme therefore set up the developments of data engineering and analytics and provided education in designing and planning industry grade solutions for big data.

The second school was Ecole Polytech for their graduate degree in innovation and management programme. The additional notes indicated on what the programme was

based, this included providing the students with the necessary skills for the application of electronics to economics of connected objects, the programme was intended for future entrepreneurs and IT consultants with an interest in the internet of things, as well as to investigate the different sectors within the field.

The third school was the University of Zurich for their bachelors' course in business, economics and informatics. The additional notes gave an overview of the course, this included students gaining knowledge from extremely large data sets which could transform and accelerate business developments. The course focussed on large data management, data management, efficient processing and analysis, as well as interactive visualisation of data sets.

The last school from Europe was the University of Copenhagen for its summer course in big data analytics. The additional notes included an overview of the course which comprised of the students learning the tools and methods necessary for large-scale data analysis based on cutting-edge research and extensive research. The course therefore focussed on big data analysis, which included a background in statistics and conventional data analysis. The course also required the student to have prior data analysis experience.

From the above data collected from the institutions based on their additional notes, it is important to note that additional notes are important for the students to understand the module better. This understanding may include what the course requires of the student, any prerequisites the student needs before enrolling in the course, as well as a general overview of the course so that the student knows exactly what to expect from the course in terms of skills and knowledge.

### 5.3 Conclusion

The data collected from the identified target group of institutions were thus collected by means of the data collection instrument (see Appendix A) that has been presented and were analysed in the best and most efficient manner to enable the mapping of skills being developed in the selected curricula with the skills that were identified in Chapter 2 of this study. The skills referred to in Chapter 2, were then mapped to the curriculum identified in terms of the institutions' data management and big data programmes and were then analysed accordingly in terms of which skills those curricula contained.

With regard to further analysis of the curricula of each school which was studied, it is also important to note that different regions provided different important analyses of their curricula. Concerning the Library and Information Science institutions, the data illustrated that all the curricula entailed data management programmes, whereas, three of the five institutions also presented big data management programmes. It is also important to note that the curricula included core and elective modules, only two of the curricula included electives in their courses, one of which was a master's degree and the other was an undergraduate degree.

From the Library and Information Science schools, the data illustrated that four of the five institutions used, included data management programmes, whereas, three of the five institutions' curricula included big data management programmes. All the institutions analysed included core modules, whereas, only one of the institutions included core and elective modules in their data management programme.

Regarding the institutions from the United States that were analysed, the data illustrated that four of the five institutions had big data management programmes in their curricula, while the fifth school only included a data management programme in its curriculum. All the big data management curricula included core modules in their programmes, while the data management curricula included core and elective modules in their programmes.

The data pertaining to the institutions from Australia that were analysed, illustrated that three of the five institutions had big data management programmes in their curricula, while the remaining two of the five institutions included data management programmes in their curricula. It is also important to note that only one of the big data management courses only presented core modules, while the rest of the courses comprised core and elective modules for their programmes.

Regarding the institutions that were analysed from South Africa, the data reveal that all three the institutions only had data management programmes in their curricula, and no big data management programs recorded as yet. Of the three data management programmes in the South African school's curricula, all three of the courses contained core and elective modules in their programmes.

An analysis of the data on the institutions from Europe revealed that three of the five-school's had big data management programmes in their curricula, while the remaining two institutions included data management programs in their curricula. Of the three big data management programmes, all three the programmes only offered core modules, while one of the two data management programmes provided core and elective modules for their curricula.

Lastly, an analysis of data obtained from the institutions from the United Kingdom illustrate that four of the five institutions had big data management programs in their curricula, while the remining one school only offered a data management programme in its curriculum. Of the big data management programmes, all of them only included core modules for their programs, and the data management programme also only included core modules in its programme.

From the above and from all the analyses presented in this chapter, it is important to note that the fields of study from all of the chosen schools from their perspective universities were of a very broad nature, but all exhibited similarities and commonalities in terms of their data management and big data management curricula. This also reveals diversity in the nature of the data management and big data curricula within the schools, which illustrates how diverse data management and big data management can be, if it is simply applied to a specific field and adapted accordingly.

With that being said, if data management and big data management programmes are being adapted to suit a specific field of study, the skills of the data and big data curator also need to be adapted to the field of study as the curator will also need knowledge of how the specific field works.

Even though there are a diverse number of fields of study into which data management and big data management can fit and be moulded, it is important to note that for the purpose of this study, the main primary focus is on the data and big data management curricula. The purpose of this study is to identify the modules, which exist in the specific curricula so that skills can be created or adapted accordingly.

Along with the field of study, modules and data management and big data management curricula, it is also important to note the level of study on which these modules and curricula exist and how they vary according to the different modules in the curricula as seen in sections

3.1 to 3.7. Another interesting fact gathered from the data from the specific schools is which the programmes that existed in the different regions, as some regions may have more big data management programmes at certain levels, while other regions may offer data management programmes at certain levels of study only.

As an analysis of skills has been conducted, it is important to note that some institutions included many skills in their curricula, whereas, other institutions had some gaps in terms of the skills being addressed in their curricula and how they mapped according to the identified skills in Chapter 2 of this study. In order to adapt a big data management curriculum in Chapter 6, it is necessary to take note of the skills which were mapped to the skills identified in Chapter 3, as well as which skills did not. This will also assist in identifying gaps in the skills and can also assist in determining which skills will be necessary to fill this gap in the newly adapted curriculum.

# Chapter 6

## 6. Recommendations

### 6.1 Introduction

This concluding chapter culminates in a recommendation of a curriculum that can be used to educate big data stewards. The chapter also focusses on important aspects of the study, such as the research questions and the associated findings. Because so little evidence could be found that big data stewards are being trained at academic institutions, the chapter includes recommendations for further study, thereby, ensuring the possible continuation of this study. Accordingly, this conclusion reflects on the study, the importance of the study and the value that can be gained from further postgraduate research in big data stewardship.

### 6.2 Research questions and most important findings

The research questions for this study were used to guide the study in the right direction and to inform the study with purpose and meaning.

The main research question of this study asked **what would a training programme for big data stewards entail?** This question is answered in section 6.4 below where the suggested curriculum content is provided in some detail.

The sub-questions consisted of five main categories. This research answered all the sub-questions successfully.

The first main category of the sub-questions asked: **what is the existing big data framework or context for data stewards?** The key finding was that there is no shortage of information about the concepts and there is ample detail about the context in which big data become a knowledge asset. This part of the research addressed issues such as the key concepts with which a big data steward should be familiar, (refer to Chapter 2, section 2.2 and Chapter 3, section 3.3), the skills, competencies, and outcomes of big data and big data stewardship (see section 3.3), a definition and the characteristics of big data (see section 2.2) and big data as the foundation of valuable information and a knowledge asset – a format to be managed and stewarded (see section 2.2.3).

The second category of sub-questions: **how does big data stewardship differ from long tail data stewardship?** This question was answered in Chapter 2, section 2.2.2, where it was reported that the differences between big data stewardship and long tail data stewardship is mainly linked to the sheer size of the data file. It was also found that big data are more likely to be centrally curated at the point of collection, while long tail data are usually curated where it makes most sense to do so. Big data are usually not accessible via a repository, while long tail data are found either in discipline or institutional repositories.

The next category of sub-questions asked: **how does data stewardship fit into the big data life cycle?** Here the focus was on the big data life cycles that are currently being promoted (see section 2.7). The characteristics of the big data life cycle models (see sections 2.7.1 and 2.7.3), the big data cycle models used in business versus big data cycle models used in research (section 2.7.2), whether the DCC curation model catered also for big data, the UK Data Archive model and big data, as well as a data stewardship intervention model for big data (refer to sections 2.7.4, 2.7.5 and 2.7.6). It was established that many life cycle models exist, and that it is difficult to indicate whether one big data life cycle is correct, and another is not. However, a big data life cycle that addressed similar components from different big data life cycles was discussed in section 2.7.1. At the end of this section, a table was created which included all the life cycle model activities deduced from all the life cycle models discussed in section 2.7. This table also included stewardship activities that assisted with understanding how big data stewardship fits into and has a role in the big data life cycle.

Subsequently, it was then necessary to ask: **what are the known challenges of big data stewardship training?** Chapter 3, section 3.2.3 addressed the challenges of big data stewardship in some detail. The most important challenges are the heterogeneity of big data, the inconsistency and incompleteness of big data, the large scale of big data, the privacy and ownership issue of big data, as well as the visualisation and collaboration of big data from a human perspective. These challenges are discussed in more detail in section 3.2.3.

The focus then shifted to the data steward. The questions explored **what the roles and responsibilities of big data stewards are.** The roles and responsibilities of big data stewards were documented in section 3.2. This section also considered how aspects such as the FAIR principles, data ownership and intellectual property rights, data governance, data management training, managing access to data, data quality control, and risk management,

influence the skills required from a data steward. To address this question further, the desired skill set, for big data stewards, was identified (refer to sections 3.4 to 3.6). Nineteen different technical skills, six personal (soft) skills and several knowledge components, required for future work environments, were identified. A competency matrix with knowledge, skills and experience required, was developed (see section 3.7).

Research was then conducted to address the last set of sub-questions. Here the author paid attention to **the big data stewardship training curriculum as addressed at a selected group of universities (international as well as South African).** The results from the research were documented in detail in Chapter 5. The data were collected by means of desktop research, and content analysis was used as the research method. The contents of both the big data and the long tail data management curricula were interrogated. The curricula of any data-related electives were included in the investigation. The field of study (see section 5.2.2), the module of data-related study (academic level), the electives of the curricula (see section 5.2.9), the skills being addressed, as well as the learning outcomes were all documented (see section 5.2.8 and attachment 6). All the prescribed works were listed, and any other facts of interest were noted. All these data were captured with the help of a standardised data collection template (see Attachment 1).

The roles, responsibilities and skills, that were found in the literature study, were used to review the content identified in the published curricula. It is important to highlight the University of Illinois – Urbana Champaign, as they were the only university (in the selected sample) that explicitly addressed data curation as a specialist training option. All the module content was provided online, while the other universities only a few presented relevant data curation-related module topics or courses that they offered, but, nowhere else was the main focus on data curation (whether big or long tail).

Based on the identification and analysis of the learning outcomes for each of the relevant learning units (module, topic or course) it was possible to map and deduce the skills that would need to be included in a curriculum that prepares individuals for the data curation responsibility. This suggested curriculum is provided in section 6.4 below.

## 6.3 General recommendations

The big data stewardship responsibilities that were identified in Chapter 3, section 3.2 (refer also to Table 10), were reviewed and extended to include the information gathered from the training curricula. The responsibilities that were reflected in Table 5.2, in Chapter 5, have now been augmented to provide a single integrated set of roles and responsibilities. For ease of reading, please note that the role on the left is directly linked to the responsibility on the right of the table. These data steward's roles were identified by Loshin (2009). Please see Table 6.1 below [refer to p48 in the consolidated document]:

*Table 6.1: Updated list of big data steward roles, responsibilities, skills and outcomes*

| **Roles of the data steward**<br><br>**(refer to 3.2)** | **Responsibilities**<br><br>**(refer to 3.2)** | **Skills necessary to address the responsibility identified (refer to 3.3 – 3.6)** | **Suggested learning outcomes for a big data curation programme** |
|---|---|---|---|
| Champion | • Facilitating and managing the appropriate and efficient use of the big data asset.<br>• Ensuring that data is FAIR (findable, accessible, interoperable and reusable) by negotiating for the infrastructure, documentation and resources to do so. | (S) Independent worker<br><br>(S) Community based data skills<br><br>(S) Building relationships | Make use of a conceptual framework to recognise the potential of data.<br><br>Influencing peers and superiors to understand the need to build the essential infrastructure.<br><br>Planning and writing essential documentation such as infrastructure funding requests, policy and guidelines for the initiative. |
| Collaborator | • Working collaboratively in teams<br>• Building relationships with other data stewards<br>• Collaborating with other data stewards to establish data standards where these may not exist | (T) Administrative skills<br><br>(T) Project management skills<br><br>(S) Building relationships | Understand the role institutions, agencies, policies, and laws play in big data curation.<br><br>Understand hierarchies and standards for data transformation and transcoding.<br><br>Understand the need for a collaborative working model - |

| Roles of the data steward (refer to 3.2) | Responsibilities (refer to 3.2) | Skills necessary to address the responsibility identified (refer to 3.3 – 3.6) | Suggested learning outcomes for a big data curation programme |
|---|---|---|---|
| | | | rather than 'we do it all ourselves' model.<br><br>Understand the responsibilities associated with a collaborative working model.<br><br>A very good understanding of the ethical use of big data. |
| Technical expert | • Interpreting heterogeneity in big data.<br>• Managing inconsistency and incompleteness in the data set.<br>• Being able to address issues of scale in the big data set.<br>• Making use of real-time techniques to keep up to date with the ever-growing amounts of data so that data overload does not occur. | (T) Administrative skills<br><br>(T) Developing policy & procedural documents<br><br>(T) Appraisal skills<br><br>(S) Building relationships<br><br>(T) Licensing of data skills<br><br>(T) Archiving skills<br><br>(T) Metadata skills | Understand and gain knowledge on data intensive computing.<br><br>Understand high performance computing.<br><br>Understand distributed computing<br><br>Understand data science constraints (i.e. inconsistencies, incompleteness and overload). |

| Roles of the data steward (refer to 3.2) | Responsibilities (refer to 3.2) | Skills necessary to address the responsibility identified (refer to 3.3 – 3.6) | Suggested learning outcomes for a big data curation programme |
|---|---|---|---|
| | • Managing the privacy and ownership of big data without compromising the value of the data.<br>• Finding ways to visualise data so that the data is useful from the human perspective.<br>• Organising and describing data (skills to illustrate understanding of the data and the purpose of the data). | -(T) Project management skills<br>(T) Data usage skills<br>(T) Data research skills<br>(T) Working with and formatting data.<br>(T) Managing data storage<br>(T) Repository skills<br>(T) Organisation skills<br>(T) Access, sharing, dissemination skills<br>(T) Practical and technical data skills | Understand the differences and similarities between advanced tools and methods in data science.<br>Understand the management of heterogeneity in data management, including schema matching techniques.<br>Understand the use and purpose of metadata, controlled vocabularies, ontologies and metadata schemas.<br>Understanding the nature of managing big data curation projects. |

| Roles of the data steward (refer to 3.2) | Responsibilities (refer to 3.2) | Skills necessary to address the responsibility identified (refer to 3.3 – 3.6) | Suggested learning outcomes for a big data curation programme |
|---|---|---|---|
| | | | Practical / hands-on experience in visualising big data.<br><br>Practical / hands-on experience in splitting big data sets into smaller subsets.<br><br>Practical / hands-on experience in merging data sets.<br><br>Practical / hands-on experience in working with metadata.<br><br>Practical / hands-on experience in using Hadoop, Spark, R and/or Python.<br><br>Practical / hands-on experience in using the cloud computing environment. |

| **Roles of the data steward**<br><br>**(refer to 3.2)** | **Responsibilities**<br><br>**(refer to 3.2)** | **Skills necessary to address the responsibility identified (refer to 3.3 – 3.6)** | **Suggested learning outcomes for a big data curation programme** |
|---|---|---|---|
| | | | Perform data derivation. |
| Knowledgeable expert | Maintaining expert level knowledge about:<br><br>• big data life cycles<br>• funder requirements<br>• the value of big data<br>• discipline knowledge, including discipline-specific methodologies<br>• Safety, security licensing and copyright of data<br>• Developing a good understanding of research ethics<br>• Understand the importance of data as a secondary resource | (K) Data lifecycles<br><br>(K) Research ethics<br><br>(K) Funder requirements<br><br>(K) Value of data<br><br>(K) Secondary data use<br><br>(S) Community based data skills<br><br>(T) Data citation skills | Understand the need for continuous learning.<br><br>Understand the broader context of data management and research data management.<br><br>Describe common data behaviours of managers, programmers, scientists and other users.<br><br>Evaluate data to understand quality in research data.<br><br>Evaluate ethical data collection practices. |

| **Roles of the data steward**<br><br>**(refer to 3.2)** | **Responsibilities**<br><br>**(refer to 3.2)** | **Skills necessary to address the responsibility identified (refer to 3.3 – 3.6)** | **Suggested learning outcomes for a big data curation programme** |
|---|---|---|---|
| | | | Evaluate ethical data use practices. |
| Trainer | Facilitation, data presentation, understanding of learning styles, and working with different types of people<br><br>Training colleagues and researchers on the following topics:<br><br>• naming conventions,<br>• versioning,<br>• labelling<br>• writing data management plans<br>• writing data documentation<br>• allocating metadata | (S) Time management<br><br>(T) Writing DMPs<br><br>(T) Administrative skills<br><br>(T) Developing policy & procedural documents<br><br>(T) Appraisal skills<br><br>(T) Licensing of data skills<br><br>(T) Archiving skills<br><br>(T) Metadata skills<br><br>(T) Citation of data skills<br><br>(T) Project management skills | Demonstrate training skills.<br><br>Demonstrate expertise in using applications and tools.<br><br>Write training material.<br><br>Explain concepts, context and practical requirements of big data management. |

| Roles of the data steward (refer to 3.2) | Responsibilities (refer to 3.2) | Skills necessary to address the responsibility identified (refer to 3.3 – 3.6) | Suggested learning outcomes for a big data curation programme |
|---|---|---|---|
| | | (T) Data usage skills<br>(T) Data research skills<br>(T) Working with and formatting<br>(T) Managing data storage<br>(T) Repository skills<br>(T) Organisation skills<br>(T) Access, sharing, dissemination skills<br>(T) Practical and technical data skills<br>(K) Data life cycles<br>(K) Funder requirements<br>(K) Value of data assets<br>(K) Discipline research requirements / methodologies | |

| Roles of the data steward<br><br>(refer to 3.2) | Responsibilities<br><br>(refer to 3.2) | Skills necessary to address the responsibility identified (refer to 3.3 – 3.6) | Suggested learning outcomes for a big data curation programme |
|---|---|---|---|
| | | (K) Safety & security<br><br>(K) Licensing and copyright<br><br>(K) Research ethics<br><br>(K) Secondary data use | |

| **Roles of the data steward**<br><br>**(refer to 3.2)** | **Responsibilities**<br><br>**(refer to 3.2)** | **Skills necessary to address the responsibility identified (refer to 3.3 – 3.6)** | **Suggested learning outcomes for a big data curation programme** |
|---|---|---|---|
| Quality assurer | Developing and promoting data quality standards<br><br>Ensuring that the data is associated with the correct metadata<br><br>Ensuring availability of the data and ensuring accessibility to the data<br><br>Considering the timeliness of the data<br><br>Checking the authorisation of the data<br><br>Ensuring **usability** (which means documentation of the data, credibility of the data, and the data's metadata)<br><br>Taking responsibility for the **reliability** of (trust in) of the data (which includes accuracy of the data, integrity of the data, consistency mechanisms of the | (T) Appraisal skills<br><br>(T) Data research skills<br><br>(T) Access, sharing, dissemination skills<br><br>(T) Practical and technical data skills | Understand the different mechanisms used for controlling quality and accessibility.<br><br>Develop documentation for governance and quality control.<br><br>Develop checklists for auditing.<br><br>Identify and use data quality standards.<br><br>Conduct quality audits.<br><br>Conduct governance audits.<br><br>Conduct trusted repository audits. |

| Roles of the data steward (refer to 3.2) | Responsibilities (refer to 3.2) | Skills necessary to address the responsibility identified (refer to 3.3 – 3.6) | Suggested learning outcomes for a big data curation programme |
|---|---|---|---|
| | data, completeness of the data, and auditability of the data)<br><br>Taking care of **relevance** issues (which includes the degree of correlation between data and content, and the data users' expectations of the given data, being able to advise on the adaptability of the data to the specific need of the user<br><br>Providing a valid description method for the given data, (which then allows for full understanding of the data by users) | | |

| **Roles of the data steward (refer to 3.2)** | **Responsibilities (refer to 3.2)** | **Skills necessary to address the responsibility identified (refer to 3.3 – 3.6)** | **Suggested learning outcomes for a big data curation programme** |
|---|---|---|---|
| Gatekeeper | • Facilitating and managing the appropriate and efficient use of the big data<br>• Managing access to data<br>• Establishing and maintaining repositories<br>• Controlling data security<br>• Ensuring the persistent identification of data sets<br>• Measuring the impact of data sets | (T) Developing policy & procedural documents<br>(T) Appraisal skills<br>(T) Licensing of data skills<br>(T) Metadata skills<br>(T) Repository skills<br>(T) Access, sharing, dissemination skills | Understand the importance of accessibility, authorization and data security.<br>Understand FAIR principles and when these cannot be used.<br>Understand data licencing.<br>Understand persistence in the identification of data objects. |
| Access provider | • Providing machine actionable landing pages enable the | (S) Community based data skills<br>(T) Administrative skills<br>(T) Developing policy & procedural documents<br>(T) Licensing of data skills | Understand data access mechanisms.<br>Develop access policies.<br>Develop machine actionable landing pages. |

| Roles of the data steward (refer to 3.2) | Responsibilities (refer to 3.2) | Skills necessary to address the responsibility identified (refer to 3.3 – 3.6) | Suggested learning outcomes for a big data curation programme |
|---|---|---|---|
| | accessing of metadata and data via query re-execution.<br>• Considering query uniqueness<br>• Ensuring that sorting of data sets is unambiguous and reproducible<br>• Setting up a standardised result set verification system<br>• Timestamping queries (when they are resolved compared to when they were made)<br>• Query identification<br>• Storing query metadata<br>• Creating automated citation texts | (T) Metadata skills<br>(T) Citation of data skills<br>(T) Data usage skills<br>(T) Working with and formatting<br>(T) Repository skills<br>(T) Access, sharing, dissemination skills | Create automated citation texts.<br>Trouble shoot access problems. |

| Roles of the data steward (refer to 3.2) | Responsibilities (refer to 3.2) | Skills necessary to address the responsibility identified (refer to 3.3 – 3.6) | Suggested learning outcomes for a big data curation programme |
|---|---|---|---|
| Ownership arbitrator | • Shaping the data policy<br>• Developing access procedures and processes<br>• Develop documentation that guide the institution in the creation, collection and consumption of the data<br>• Licensing data<br>• Guiding the proper citation of data | (T) Administrative skills<br>(T) Developing policy & procedural documents<br>(T) Appraisal skills<br>(T) Licensing of data skills<br>(T) Citation of data skills<br>(T) Data usage skills | Understand data ownership issues.<br>Understand the implications of ownership models. |
| Data migrator | • Making modifications to the data infrastructure (which includes technology migration for new data representation)<br>• Migration verification (which includes ensuring that queries can be re-executed correctly for the | (T) Administrative skills<br>(T) Developing policy & procedural documents<br>(T) Project management skills<br>(T) Data usage skills<br>(T) Practical and technical data skills | Understand the implications of moving data from one platform to another.<br>Understand the possible migration errors. |

| Roles of the data steward (refer to 3.2) | Responsibilities (refer to 3.2) | Skills necessary to address the responsibility identified (refer to 3.3 – 3.6) | Suggested learning outcomes for a big data curation programme |
|---|---|---|---|
| | verification of successful data and query migration) | | |
| Preservation manager | Archiving data<br><br>Managing data storage | (T) Administrative skills<br><br>(T) Developing policy & procedural documents<br><br>(T) Appraisal skills<br><br>(T) Archiving skills<br><br>(T) Metadata skills<br><br>(T) Project management skills<br><br>(T) Working with and formatting<br><br>(T) Managing data storage<br><br>(T) Repository skills<br><br>(T) Organisation skills | Understand the preservation process.<br><br>Understand big data preservation challenges.<br><br>Understand big data preservation options in storage infrastructure and formats.<br><br>Create effective data storage schemas.<br><br>Trouble shoot migration errors. |

| Roles of the data steward (refer to 3.2) | Responsibilities (refer to 3.2) | Skills necessary to address the responsibility identified (refer to 3.3 – 3.6) | Suggested learning outcomes for a big data curation programme |
|---|---|---|---|
| | | (T) Practical and technical data skills | |
| Project manager | • Project planning and management<br>• Communicating new and changed business requirements to individuals affected<br>• Managing time and deadlines<br>• Working independently<br>• Developing budgets | (S) Time management<br>(S) Independent worker<br>(S) Attention to detail<br>(S) Building relationships<br>(S) Developing budgets<br>(T) Project management skills | Understand the challenges and opportunities associated with people management.<br>Understand the challenges and opportunities associated with financial management.<br>Understand the challenges and opportunities associated with |

| Roles of the data steward (refer to 3.2) | Responsibilities (refer to 3.2) | Skills necessary to address the responsibility identified (refer to 3.3 – 3.6) | Suggested learning outcomes for a big data curation programme |
|---|---|---|---|
| | | | big data curation infrastructure management.<br><br>Working knowledge of project management systems.<br><br>Working knowledge of scheduling.<br><br>Working knowledge of milestones and deliverables. |

*Table 13: Updated list of big data steward roles, responsibilities, skills and outcomes*

**Legend**

| Skills necessary to address the responsibility identified | |
|---|---|
| (S) – Soft skills | (K) – Knowledge component |
| (T) – Technical skills | |

Table 6.1hows the responsibilities as well as the skills required to conduct big data curation. These two aspects then drive the learning outcomes – which is reflected in the fourth column. Table 6.1, above, is an integration of the roles, responsibilities and skills identified in literature (see Chapter 3), the detail collected from the data analysis (see Chapter 5), the outcomes identified in Chapter 5 and an interpretation of necessary outcomes from an evaluation of the integrated list of responsibilities associated with the curation roles.

## 6.4 Recommendations – designing the big data training curriculum

It is recommended that the responsibilities associated with data stewardship should drive the training effort. In Table 6.2 below, the curation learning outcomes were mapped according to the academic level which the outcome originated during the data collection process. These academic levels are also stated in Appendix 3. It was seen as important to keep these learning outcomes within their given academic levels for accuracy of the research.

*Table 6.2: : Learning outcomes paired to appropriate academic level*

| **Curation learning outcomes** | **Academic level from which this outcome originated** |
|---|---|
| • Understand the nature of managing big data curation projects (interpretation outcome)<br>• Understand the challenges and opportunities associated with big data curation infrastructure management (interpretation outcome)<br>• Understand and gain knowledge on data intensive computing<br>• Understand high performance computing<br>• Understand distributed computing<br>• Explain concepts, context and practical requirements of big data management (interpretation outcome).<br>• Make use of conceptual frameworks to recognise the potential of data | Certification |
| • Understand data science constraints (that is, inconsistencies, incompleteness and overload) (interpretation outcome)<br>• Practical/hands-on experience in using Hadoop, Spark, R and / or Python (interpretation outcome).<br>• Practical/hands-on experience in using the cloud computing environment (interpretation outcome)<br>• Practical / hands-on experience in visualising big data (interpretation outcome) | Undergraduate |

| Curation learning outcomes | Academic level from which this outcome originated |
|---|---|
| • Practical / hands-on experience in splitting big data sets into smaller subsets (interpretation outcome)<br>• Practical / hands-on experience in merging data sets (interpretation outcome) | |
| • Understand the importance of accessibility, authorisation and data security (interpretation outcome)<br>• Understand FAIR principles and when these cannot be used (interpretation outcome)<br>• Understand data licensing (interpretation outcome)<br>• Understand data ownership issues (interpretation outcome)<br>• Understand the implications of ownership models (interpretation outcome)<br>• Understand the implications of moving data from one platform to another (interpretation outcome)<br>• Understand hierarchies and standards for data transformation and transcoding<br>• Understand the preservation process (interpretation outcome)<br>• Understand big data preservation challenges (interpretation outcome)<br>• Understand big data preservation options in storage infrastructure and formats (interpretation outcome)<br>• Understand persistence in the identification of data objects (interpretation outcome) | Graduate |

| Curation learning outcomes | Academic level from which this outcome originated |
|---|---|
| • Understand the management of heterogeneity in data management, including schema matching techniques<br>• Understand the use and purpose of metadata, controlled vocabularies, ontologies and metadata schemas (interpretation outcome)<br>• Understand the role of institutions, agencies, policies, and laws play in data curation (interpretation outcome)<br>• Identify and use data quality standards (interpretation outcome)<br>• Perform data derivation<br>• Create automated citation texts (interpretation outcome)<br>• Develop machine actionable landing pages (interpretation outcome)<br>• Practical / hands-on experience in working with metadata (interpretation outcome)<br>• **Describe common data behaviours of managers, programmers, scientists, and other users** | |
| • Understand the need for continuous learning (interpretation outcome).<br>• Understand the challenges and opportunities associated with people management (interpretation outcome).<br>• Understand the challenges and opportunities associated with financial management (interpretation outcome). | Postgraduate |

| Curation learning outcomes | Academic level from which this outcome originated |
|---|---|
| • Understand the responsibilities associated with a collaborative working model (interpretation outcome).<br>• Understand the need for a collaborative working model – rather than 'we do it all ourselves' model (interpretation outcome).<br>• Planning and working essential documentation such as infrastructure funding requests, policy and guidelines for the initiative (interpretation outcome).<br>• Write training material (interpretation outcome).<br>• Influencing peers and superiors to understand the need to build the essential infrastructure (interpretation outcome).<br>• Demonstrate training skills (interpretation outcome).<br>• Working knowledge of milestones and deliverables (interpretation outcome).<br>• Working knowledge of scheduling (interpretation outcome).<br>• Working knowledge of project management systems (interpretation outcome). | |
| • Understand the different mechanisms used for controlling quality and accessibility (interpretation outcome).<br>• Understand data access mechanisms (interpretation outcome).<br>• Understand the possible migration errors (interpretation outcome).<br>• Develop access policies (interpretation outcome). | Honours |

| Curation learning outcomes | Academic level from which this outcome originated |
|---|---|
| • Create effective data storage schemas.<br>• Troubleshoot access problems (interpretation outcome).<br>• Troubleshoot migration errors (interpretation outcome). | |
| • Understand the differences and similarities between advanced tools and methods in data science (interpretation outcome).<br>• An extremely good understanding of the ethical use of big data (interpretation outcome).<br>• Understand the broader context of data management and research data management (interpretation outcome).<br>• Demonstrate expertise in using applications and tools (interpretation outcome).<br>• Evaluate data to understand quality in research data (interpretation outcome).<br>• Evaluate ethical data use practices (interpretation outcome).<br>• Evaluate ethical data collection practices (interpretation outcome).<br>• Develop documentation for governance and quality control (interpretation outcome).<br>• Develop checklists for auditing (interpretation outcome).<br>• Conduct quality audits (interpretation outcome).<br>• Conduct governance audits (interpretation outcome).<br>• Conduct trusted repository audits (interpretation outcome). | Master's |

## Legend

| **South African qualifications translated into International qualifications** |
|---|
| Certificate: Anything from 1 week to 6 months to a one year, depending on the content and difficulty of programme. |
| Undergraduate: t3 years' study in the South African context, four years' study in the international context (inclusion of honours degree) |
| Honours: 1 to 2 years of study. Only applicable to South African degrees. It is the fourth year of study and the year between undergraduate and postgraduate studies. |
| Postgraduate |
| Master's: 1 to 2 years' study. Can be either course work and research based, or only research based. |
| Doctorate: 2 + years' study. Research based degree. |

(Source: Types of different degree levels. https://study.com/different_degrees.html)

In Table 6.2 above, one can see the allocation of the different curation learning outcomes to the different academic levels, which are the academic levels the researcher identified in the relevant sources during the data collection period of this research study.

For future use of the learning outcomes, wider than this research, it was decided that the learning outcomes should be called topics. The reason for this is because different information science schools may have different programmes and structures with regard to their programmes. Theses learning outcomes, therefore, can either be used as topics within a specific module of a programme, or the topic can be expanded into a module within a specific degree or programme. The above table is simply a guide regarding which topic could be addressed at which academic level.

With regard to adding credit values to each topic, a restriction would be that each university is different in terms of the system they use to award a certain number of credits. Each department or faculty in a specific university, may also use different systems. Therefore, it will be difficult to achieve consistency with regard to awarding credits and that is why no credits were suggested, as the awarding of a certain number of credits will depend on the system of a specific university and faculty.

It is also important to note that the remaining learning outcomes which were not considered as curation learning outcomes in Appendix 3, can still be deemed to be relevant for this research. As this research is intended for an information science programme for the big data curator, it is not necessary to discard other departments who would perhaps like to take these topics as electives for their studies. This can also apply to the information science programme students as they could perhaps incorporate the remaining learning outcomes that do not apply to the information science programme exactly, as electives.

These electives have thus been identified in Appendix 3 and have been provided in the table below for ease of reference. Please note that the table below will indicate the learning outcome, as well as the field of study from which it comes and to which it belongs. This field of study will then also indicate that information science and data related modules are and can be prevalent in many other fields other than that of data science and information science.

*Table 6.3: Learning outcomes and field of study*

| Learning outcome to be taken as elective | Academic field of origin |
|---|---|
| **Certification topics** | |
| • Be able to design methodologies to develop big data solutions.<br>• Gain knowledge and experience of machine learning.<br>• Gain knowledge and experience of artificial Intelligence.<br>• Gain knowledge and experience of networking.<br>• Learn how to use high performance computing facilities.<br>• Address big data issues by using distributed systems and big data.<br>• Use super computers and high-performance computing clouds.<br>• Understand and gain knowledge of programming and local remote visualisation techniques. | Computing Science |
| • Be able to interpret data for strategic decision-making.<br>• Be able to use analytical frameworks for marketing strategies.<br>• Use algorithmic tools for digital and non-digital marketing goals.<br>• Recognise big data within a firm's marketing strategy. | Marketing |
| • Set up basic big data analysis.<br>• Become acquainted with tools such as data cleaning, statistical methods for large datasets, data stream analysis, finding patterns and outliers in big data.<br>• No data curation outcomes. | Statistics |
| **Undergraduate topics** | |
| • Contribute to the field of Computer Science. | Computer science |

| Learning outcome to be taken as elective | Academic field of origin |
|---|---|
| • Gain practical knowledge and skills from the course research project.<br>• Conduct business planning and programming projects. | |
| • Compare data streaming methods such as sampling, sketching and hashing.<br>• Apply spatial data methods.<br>• Apply large scale graphs, vectors and document processing methods.<br>• Evaluate the suitability of different distributed technologies for big data processing | Information technology |
| • Explain how the statistical features of big data impact traditional statistical methods and theory.<br>• Discuss the random matrix theory and its application in statistics on large scale.<br>• Summarise the theory of sequential prediction and management of streaming data.<br>• Demonstrate the use of computational tools to work with big and streaming data sets. | Business and economics |
| • Understand and be able to apply data visualisation techniques.<br>• Understand and be able to apply statistical analysis.<br>• Understand and be able to apply machine learning methods.<br>• Interpret results of data analysis.<br>• Solve problems with analytical techniques.<br>• Understand and be able to apply Python, METLAB.<br>• Gain practical presentation skills. | Engineering |
| • Be able to apply data analysis. | Political science |

| Learning outcome to be taken as elective | Academic field of origin |
|---|---|
| • **Honours topics** | |
| • Gain knowledge on all facets of big data analytics.<br>• Apply and gain knowledge on machine learning and optimisation of statistics.<br>• Understand the introduction to big data analytics. | Computer science and applied mathematics |
| • Understand importance of data for managing organisations through information systems.<br>• Be able to perform data manipulation<br>• Apply knowledge to real-world situations by data gathering and processing of big data | Information technology and business management |
| • **Postgraduate topics** | |
| • Gain knowledge of non-traditional and large-scale data applicators.<br>• Gian knowledge of properties of social networking.<br>• Gain knowledge of the fundamentals of NoSQL systems.<br>• Gain experience with NoSQL systems and Hadoop.<br>• Work and gain knowledge of database management systems.<br>• Work and gain knowledge of database design principles.<br>• Work and gain knowledge of database design tools. | Information and library science |
| • Become familiar with industry trends in big data systems.<br>• Understand the trade of designing and implementing big data systems.<br>• Become more knowledgeable in decision making for big data scenarios.<br>• Develop research skills. | Engineering and applied science |

| Learning outcome to be taken as elective | Academic field of origin |
|---|---|
| • Perform programming, debugging, and performance profiling. | |
| • Apply big data analytics techniques.<br>• Understand data infrastructures used for data analytics.<br>• Gain experience working with big data analytics, data science, and big data for understanding.<br>• Gain knowledge on big data management infrastructure. | Data science |
| • **Master's topics** | |
| • Be able to apply big data science techniques to the organisation<br>• Learn how to avoid pitfalls in big data analytics.<br>• Deploy machine learning algorithms to mine data.<br>• Be able to interpret analytical models for effective decision making<br>• Convert datasets into models through predictive analytics.<br>• Recognise challenges with big data algorithms.<br>• Learn how to effectively represent data. | Data, systems, and society |
| • Understand machine learning<br>• Adapt to the evolution of data science. | Data science |
| • Apply big data science in different domains.<br>• Understand machine and statistical learning.<br>• Gain knowledge on mathematical optimisation for big data science.<br>• Understand architects available for processing big data.<br>• Understand the different research methods for big data science. | Information technology |

| Learning outcome to be taken as elective | Academic field of origin |
|---|---|
| • Gain practical experience working with big data life cycle.<br>• Gain research-based big data experience. | |

**Legend**

| South African qualifications translated to International qualifications |
|---|
| Certificate: Anything from 2 weeks to 6 months to a 1 year, depending on content and difficulty of programme. |
| Undergraduate: 3 years of study in South African context. 4 years of study in international context (inclusion of honours degree). |
| Honours: 1 to 2 years of study. Only applicable to South African degrees. It is the fourth year of study and the year between undergraduate and postgraduate studies. |
| Postgraduate |
| Masters: 1 to 2 years of study. Can be either course work and research based, or only research based. |
| Doctorate: 2 + years of study. Research based degree. |

(Source: Types of Different Degree Levels, https://study.com/different_degrees.html)

An important aspect to note is that the curation topics presented in Table 3.2 are not meant for or aimed at attracting the typical academic librarian responsible for the training of learning-focussed support. The topics presented above are intended to assist with the big data curation skills needed with regard to the information professional who is seeking the necessary technical skills, data skills, as well as analytical skills to support research. However, this does not preclude the librarian from taking the topics, it is just not aimed at the librarian's skill development and career.

Ultimately, the topics have been selected to develop big data stewards. These topics are aimed at the knowledge and skill development of the big data steward and aim at assisting stewards with their designated roles and responsibilities.

## 6.5 Recommendations for further research

The research conducted for this study can lead to further doctoral research, as well as for other master's research studies. Other than that, the main aspects for future research could be the following:

6.5.1 How can this research be adapted to different environments? For example, in a South African context, can the topics be adapted to a specialisation field for information science or can the topics just be taken as electives for specific programmes?

6.5.2 From the topics stated above, which of them are specialisations for other departments or environments other than information science, and if they are, can they be considered as topics that are fed into a module or can they be adapted into modules themselves?

As these topics are based within the field of information science, it is also important to consider the big data curator and the skills development of the curator. It can then also be recommended that further research should be conducted on big data curation within the information science field, as it already has many direct links with the field.

Furthermore, within this study, topics were created according to the learning outcomes which were based on the analysis of skills, different curricula and different learning outcomes from different universities and departments across the globe. It may now be taken a step further in terms of creating a full curriculum based on the topics presented in this research. If the topics and analysis on the data collected within the research can be taken a step further, a full curriculum from first year level through to doctoral level can be created, which addresses the skills needed for big data stewardship.

## 6.6 Concluding remarks

In summary, the purpose of this research was to develop a well-researched curriculum for big data stewardship training that could be used by a prominent South African academic institution. It was also aimed at gaining a clear understanding of the framework and/or context for big data, to define the role that data stewards could play within a generic big data life cycle, to document the known challenges of big data stewardship, and to establish how institutions of higher learning (international as well as South African universities) are approaching the challenge to train big data stewards. All these objectives were met successfully

The skills needed for big data stewardship were identified and reported in the literature review and have played an important role in the mapping of learning outcomes as these are needed to address specific skill-requirements. The learning outcomes, which are now more

appropriately called topics, are based on the analysis that was conducted in Chapter 5 and have played a pivotal role in the outcome of this research study.

These outcomes serves as foundational work for the planning process as well as for further studies that need to be conducted both in South Africa but also internationally.

**References**

Ajayi, O.V. 2017. Distinguish between primary sources of data and secondary sources of data. Faculty of Education Department of Curriculum and Teaching, Makurdi. 2-5.

Alhinn, N.K., & Rababah, O. 2018. Analysis the relationship between big data and the knowledge management process: A field study in Jordanian commercial banks in Amman. Middle East University, Amman – Jordan, 1-76.

Alleman, G. 2018. Education and Certification as a Certified Data Steward. Master Data Management. [ONLINE]. Available at: https://www.masterdata.co.za/index.php/cds/data-stewardship-training. [Accessed 23 April 2018].

Alshenqeeti, H. 2014. Interviewing as a data collection method: A critical review. *English Linguistics Research.* 3(1): 39-45.

ANDS. 2018. Information Specialists and Data Librarian Skills. [online]. Available at: http://www.ands.org.au/working-with-data/data-management/overview/data-management-skills/information-specialists-and-data-librarian-skills. [Accessed: 16 August 2018].

ANDS. 2017. The FAIR data principles. [ONLINE] Available at: https://www.ands.org.au/working-with-data/fairdata. [Accessed: 7 October 2018].

Atieno, O.P. 2009. An analysis of the strengths and limitation of qualitative and quantitative research paradigms. *Problems of Education in the 21st Century*. 13:13-17.

Banuta, R. 2015. Content analysis and discourse analysis research methods. SlideShare. [online]. Available at: https://www.slideshare.net/tesono/content-analysis-and-discourse-analysis. [Accessed: 19 October 2017].

Barth, P., Bean, R., & Davenport, T.H. 2012. How 'big data' is different. *MIT Sloan Management Review.* 54(1): 22-24.

Beulke, D. 2011. Big data impacts data management: The 5 Vs of big data. [online]. Available at: https://davebeulke.com/big-data-impacts-data-management-the-five-vs-of-big-data/. [Accessed: 20 November 2019].

(Biehn, N. 2018. The missing V's in big data: Viability and value. [online]. Available at: https://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/. [Accessed: 20 November 2019].

Big data & analytics for CRM. 2018. Big data & analytics for CRM. [online]. Available at: https://www.gibs.co.za/news-events/events/open-programmes/pages/big-data--analytics-for-crm.aspx. [Accessed 23 April 2018].

Bloomberg Professional Services. 2018. 7 phases of a data life cycle | Bloomberg Professional Services. [ONLINE] Available at: https://www.bloomberg.com/professional/blog/7-phases-of-a-data-life-cycle/. [Accessed 09 February 2018].

Borgman, C.L. 2014. *Big data, little data, no data: Scholarship in the networked world*. Cambridge, Massachusetts: MIT Press.

Boyd, D., & Crawford, K. 2012. Critical questions for big data. Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society,* 15:5, 662-679, DOI: 10.1080/1369118X.2012.678878,

Braun, V., & Clarke, V. 2014. Using thematic analysis in psychology. *Qualitative Research in Psychology.* 3(2): 77-101.

Brenner, A., Lyon, L., & Mattern, E. 2016. Learning by teaching about RDM: An active learning model for internal library education. *IDCC16 Practice Paper*. 3-4.

Brikci, N., & Green, J. 2007. A guide to using qualitative research methodology. Medecins Sans Frontieres, Research Unit, London School of Hygiene and Tropical Medicine, 2-28. https://evaluation.msf.org/sites/evaluation/files/a_guide_to_using_qualitative_research_methodology.pdf

Brown, S., Bruce, R., & Kernohan, D. 2015. Directions for research data management in UK Universities. *Joint Information Systems Committee*, 7.

Brubaker, J.R., Jed R. Brubaker, Lynn S. Dombrowski, Anita M. Gilbert, Nafiri Kusumakaulika, Gillian R. Hayes. 2014. Stewarding a legacy: responsibilities and relationships in the management of post-mortem data. Proceedings of CHI. 4157-4166.

Bughin, J. 2016. Big data: getting a better read on performance. *McKinsey Quarterly*, 1-4.

Buhl, H.U., Roglinger, M., Moser, F., & Heidemann, J.2016. Big Data: A Fashionable Topic with(out) Sustainable Relevance for Research and Practice? *Springer Link.* 5(2): 65-69.

Burton, M., Lyon, L., Erdmann, C., & Tijerina, B. 2017. Shifting to data savvy, *The Future of Data Science in Libraries.* 19.

Burton, Matt and Lyon, Liz and Erdmann, Chris and Tijerina, Bonnie (2018) *Shifting to data savvy: The future of data science In:Libraries. project report*. University of Pittsburgh, Pittsburgh, PA, 19.

BusinessDictionary.com. 2019. What is responsibility? Definition and meaning - BusinessDictionary.com. [onlinE]. Available at:

http://www.businessdictionary.com/definition/responsibility.html. [Accessed: 25 November 2019].

Cai, L., & Zhu, Y. 2015. The challenges of data quality and data quality assessment in the big data are. *Data Science Journal.* 14(2): 1-10.

Castro, S. 2014. Optimising your data management for big data. *Journal of Direct, Data and Digital Marketing Practice,* 16(1): 15-18.

Cawthorne, J. 2015. Knowledge management and big data: Strange bedfellows? [online]. Available at: https://www.cmswire.com/social-business/knowledge-management-and-big-data-strange-bedfellows/. [Accessed: 6 March 2018].

Chen, M., Mao, S., & Liu, Y. 2014. Big data: A survey. *Mobile Networks and Applications*. 19(2): 171-209.

Chisholm, M. 2015. 7 phases of a data life cycle. Data Governance. [online]. Available at: https://www.bloomberg.com/professional/blog/7-phases-of-a-data-life-cycle/. [Accessed: 4 October 2018].

Chu, H. 2015. Research methods in library and information science: A content analysis. *Library and Information Science Research. Research Gate,* 37(1): 36-41.

Coates, H.L. 2014. Building data services from the ground up: Strategies and resources. *Journal of eScience Librarianship.* 3(1): 52-59.

Cockayne, D.G. 2016. Affect and value in critical examinations of the production and presumption of big data. *Big Data and Society*, 1-11. doi: 10.1177/2053951716640566.

CODATA. 2018. South Africa: Data Citation Workshop 2015 – CODATA. [online]. Available at: http://www.codata.org/task-groups/data-citation-standards-and-practices/international-

series-of-data-citation-workshops/south-africa-data-citation-workshop-2015. [Accessed: 16 August 2018].

Corea, F. 2016. Chapter 2: What data science means to the business. In: *Big Data Analytics: A* Management Perspective*.* New York: Springer, 5-17.

Creswell, J.W. 2015. *A concise introduction to mixed methods research*. Thousand Oaks, California: Sage Publications.

Crossman, A. 2017. An overview of qualitative research methods. Direct observation, interviews, participation, immersion, and focus groups. [online]. Available at: https://www.thoughtco.com/qualitative-research-methods-3026555. [Accessed: 29 January 2018].

Data Analysis | UCT Online Short Course - GetSmarter. 2018. Data analysis | UCT Online Short Course - GetSmarter. [online]. Available at: https://www.getsmarter.com/courses/za/uct-data-analysis-online-short-course. [Accessed 23 April 2018].

DataONE. 2018. Plan data management early in your project. [ONLINE]. Available at: https://www.dataone.org/best-practices/plan-data-management-early-your-project. [Accessed: 16 August 2018].

DataONE. 2019. DataONE: Data Observation Network for Earth. Data Life Cycle. [ONLINE]. Available at: https://www.dataone.org/data-life-cycle. [Accessed: 25 November 2019].

DATAVERSITY. 2018. Implementing a Data Stewardship Program - A Two Day Seminar - DATAVERSITY. [ONLINE]. Available at: http://www.dataversity.net/implementing-data-stewardship-program-two-day-seminar/. [Accessed: 23 April 2018].

Datt, S., & Datt, S. 2016. Limitations and weaknesses of qualitative research methods. Project Guru. [online]. Available at: https://www.projectguru.in/publications/limitations-qualitative-research/. [Accessed: 8 October 2017].

Davenport, T.H., *Barth, P., & Bean, R.* 2012. How 'Big Data' is Different. MIT Sloan Management Review. [ONLINE]. Available at: https://sloanreview.mit.edu/article/how-big-data-is-different/. [Accessed: 25 November 2019].

DCC (Digital Curation Centre). 2018. DCC Curation Lifecycle Model. [onlinE]. Available at: http://www.dcc.ac.uk/resources/curation-lifecycle-model. [Accessed: 16 August 018].

Debois, S. 2016. 9 Advantages and disadvantages of questionnaires. [ONLINE]. Available at: https://surveyanyplace.com/questionnaire-pros-and-cons/. [Accessed: 8 October 2017].

De Prins, S., & Veldhoen, A. 2014. Applying big data to risk management: Transforming risk management practices within the financial services industry. REPLY AVANTAGE, pp.1-16.

De Sherbinin, A., Faustmann, E., & Edmunds, R. 2018. What every early career researcher should know about research data management. World Data System. [ONLINE]. Available at: https://www.icsu-wds.org/files/what-every-ecr-should-know-about-rdm-pdf" [Accessed: 23 April 2018].

Doyle, A. 2019. What are soft skills? [ONLINE]. Available at: https://www.thebalancecareers.com/what-are-soft-skills-2060852. [Accessed: 25 November 2019].

Elo, S., & Kyngas, H. 2008. The qualitative content analysis process. *Journal of Advanced Nursing*. 62(1): 107-115.

Erl, T., Buhler, P., & Khattak, W. 2016. Big data adoption and planning considerations. *Big data fundamentals: Concepts, drivers and techniques*, 11.

Farley, A. 2019. Technical skills. Investopedia. [ONLINE]. Available at: https://www.investopedia.com/terms/t/technical-skills.asp. [Accessed: 25 November 2019].

Florentine, S. 2017. 10 critical security skills every IT team needs. CIO. [online]. Available at: https://www.cio.com/article/3228965/it-skills-training/10-critical-security-skills-every-it-team-needs.html. [Accessed: 12 October 2018].

Floyd, J., & Fowler, J. 2014. *Survey research method*s. 5th ed. Center for survey research. Boston: University of Massachusetts.

Force 11. 2016. The FAIR Data Principles. Force 11. The Future of Research Communications and e-Scholarship. [online]. Available at: https://www.force11.org/group/fairgroup/fairprinciples. [Accessed: 25 November 2019].

Fraley, R.C. & Hudson, N.W. 2014. Review of intensive longitudinal methods: An introduction to diary and experience sampling research. *The Journal of Social Psychology*. 154(1): 89-91.

Freitas, A., & Curry, E. 2016. Big data curation. *New Horizons for a Data-Driven Economy*. 6: 87-115.

Fukada, M. 2018. Nursing competency: Definition, structure and development. *YONAGO ACTA MEDICA.* 61(1): 1-7.

Genova, F. & Horstmann, W. 2016. Long tail of data. *E-IRG Task Force Report*. e-IRG secretariat AN The Hague, The Netherlands:EIRG, 1-17.

Gibson, H. 2013. Practical guidelines for starting an institutional repository (IR). Stellenbosch University Library. *ResearchGate*, 13: 7-138.

GoFair. 2019. FAIR Principles. GoFair. [ONLINE]. Available at: https://www.go-fair.org/fair-principles/. [Accessed: 25 November 2019].

Gonzalez, J., & Patten, L.G. 2017. Benefits and weaknesses of survey research. [online]. Available at: https://surveymethods.com/blog/benefits-and-weaknesses-of-survey-research/. [Accessed: 29 January 2018].

Hebbar, P. 2017. Who is a data steward and what are his roles and responsibilities? [online]. Available at: https://analyticsindiamag.com/data-steward-roles-responsibilities/. [Accessed: 19 April 2018].

Heidorn, P.B. 2008. Shedding light on the dark data in the long tail of science. *Library Trends.* 57(2).

Higgins, S. 2008. The DCC curation lifecycle model. *The International Journal of Digital Curation*. 1(3): 135-139.

Hilbert, M. 2016. Big data for development: A review of promises and challenges. *Development Policy Review.* 34(1): 135-174.

IFLA. 2018. IFLA – About the Big Data Specialist Interest Group. [ONLINE]. Available at: https://www.ifla.org/about-big-data. [Accessed: 16 August 2018].

Ismail, N. 2016. Big data in the developing world. The many crises facing developing nations can be solved as more and more people from these countries acquire mobile devices, contributing to the pool of big data. Information age. [online]. Available at: http://www.information-age.com/big-data-developing-world-123461996/. [Accessed: 23 April 2018].

ISO/IEC JTC 1. 2014. Big data Preliminary report. [online] Available at: https://www.iso.org/files/live/sites/isoorg/files/developing_standards/docs/en/big_data_report-jtc1.pdf. [Accessed: 11 April 2018].

Jagadish, H. & Gehrke, Johannes & Labrinidis, Alexandros & Papakonstantinou, Yannis & Patel, Jignesh & Ramakrishnan, Raghu & Shahabi, Cyrus. (2014). Big data and its technical challenges. *Communications of the ACM*, 57(7):86-94. 10.1145/2611567,

Jarosciak, J. 2017. The significance of big data lifecycle management (BDLM). Jozef Jarosciak Blog. [online]. Available at: https://www.joe0.com/2017/01/15/the-significance-of-big-data-lifecycle-management-bdlm/. [Accessed: 9 February 2018].

Jones, S., Pryor, G., & Whyte, A. 2013. How to develop research data management services – a guide for HEIs. DCC How to Guides. Edinburgh: Digital Curation Centre. Available at: http://www.dcc.ac.uk/resources/how-guides" http://www.dcc.ac.uk/resources/how-guides. [Accessed: 16 August 2018].

Karel, R. 2013. Metadata management for holistic data governance. *Informatica*, 1-17.

Kim, J., Moen, W.E., & Warga, E. 2013. Competencies required for digital curation: An analysis of job advertisements. *The International Journal of Digital Curation*, 8(1): 69.

Kitchin, R. 2015. Big data and official statistics: Opportunities, challenges and risks. Statistical Journal of the IAOS. 31(3): 471-481.

Kolb, J. & Kolb, J., 2013. The big data revolution. Scotts Valley, California, United States: CreateSpace Independent Publishing Platform, 10.

Kshetri, N. 2016. *Big data's big potential in developing economies: Impact on agriculture, health and environmental security*. Greensboro, North Carolina: CABI

Lagoze, C. 2014. Big data, data integrity, and the fracturing of the control zone. *Big Data and Society*, 1(2):1 -11.

Lee, C.A., Tibbo, H.R., & Schaefer, J.C. 2007. Defining what digital curators do and what they need to know: The DigCCur project. In *Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries,* JCDL 2007: Building and Sustaining the Digital Environment, Vancouver, BC, Canada, 6/18/07. DOI: 10.1145/1255175.1255183.

Leedy, P.D., & Ormrod, J.E. 2015. *Practical research: Planning and design* 11: 98-341.

Lewis, M. 2015. Qualitative inquiry and research design: choosing among five approaches. *Health Promotion Practice*. 16(4):473-475.

Loshin, D. 2009. Data governance for master data management. *Science Direct.* DOI: 10.1016/B978-0-12-374225-4.00004-7.

Loukissas, Y.A. 2016. A place for big data: Close and distant readings of accessions data from the Arnold Arboretum. *Big Data and Society*, 3(2):1–20.

Lyon, L. 2012. The informatics transform: Re-engineering libraries for the data decade. *The International Journal of Digital Curation*. 7(1):129-130.

Lyon, L., & Mattern, E. 2016. Education for real-world data science roles (Part 2): A translational approach to curriculum development. *IDCC16 Practice Paper*. 4.

Malinowski, C. 2016. Data management: File organisation. Cambridge, Massachusetts, MITLibraries.

Mamabolo, M.A., Kerrin, M., & Kele, T. 2017. Entrepreneurship management skills requirements in an emerging economy: A South African outlook. *The Southern African Journal of Entrepreneurship and Small Business Management.* 9(1):1-10.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. 2011. Big data: The next frontier for innovation, competition, and productivity. *The McKinsey Global Institute*, 1–143.

Marsh, C., Wackerman, D., & A.W. Stubbs, J.A.W. 2017. Creating an institutional repository: elements for success! *The Serials Librarian*, 72:1-4, 3-6, DOI: 10.1080/0361526X.2017.1297587.

McLeod, S. 2014. Sampling methods. SimplyPsychology. [online] Available at: https://www.simplypsychology.org/sampling.html. [Accessed: 26 July 2018].

Miemoukanda, M. 2017. The state of big data and analytics in South Africa: driving digital transformation. IDC FutureScape. [online]. Available at: https://www.idc.com/getdoc.jsp?containerId=CEMA42083917. [Accessed: 24 January 2018].

MIT in Big Data Science (Stream C). 2018. MIT in big data science (Stream C). [ONLINE]. Available at: http://www.up.ac.za/en/school-of-information-technology/article/2324622/mit-in-big-data-science-stream-c. [Accessed 23 April 2018].

Molloy, L., & Snow, K. 2012. The data management skills support initiative: Synthesising postgraduate research data management. *The International Journal of Digital Curation*. 7(2): 106-107.

Mutuku, L. 2016. The big data challenge for developing countries. [ONLINE]. Available at: https://twas.org/article/big-data-challenge-developing-countries. [Accessed: 24 April 2018].

Neuendorf, K.A. 2016. The content analysis guidebook. *Language Arts & Disciplines* 2: 45342.

Oflazoglu, S. 2017. Qualitative versus quantitative research. IntechOpen. [ONLINE]. Available at: https://www.intechopen.com/books/qualitative-versus-quantitative-research. [Accessed: 6 November 2018].

Ohri, A. 2015. Big data initiatives in developing nations. [ONLINE]. Available at: http://www.ibmbigdatahub.com/blog/big-data-initiatives-developing-nations. [Accessed: 24 April 2018].

Packt Subscription. 2019. The business analytics lifecycle. [ONLINE] Available at: https://www.packtpub.com/mapt/book/big_data_and_business_intelligence/9781783282159/4/ch04lvl1sec33/the-business-analytics-life-cycle. [Accessed 15 February 2019].

Patil, M.R., & Thia, F. 2013. *Pentaho for big data analytics*. Birmingham: PACKT.

Peters, D. 2017. Africa must keep its rich, valuable data safe from exploitation. University of Cape Town. [ONLINE]. Available at: https: www.news.uct.ac.za/article/-2017-11-22-africa-must-keep-its-rich-valuable-data-safe-from-exploitation. [Accessed: 24 April 2018].

Posa, R. 2015. Introduction to hadoop, big data life cycle management. [ONLINE]. Available at: https://www.journaldev.com/8795/introduction-to-hadoop. [Accessed: 9 February 2018].

Pouchard, L. 2015. Revisiting the data life cycle with big data curation. *International Journal of Digital Curation* 10(2): 176-192.

Pouchard, L. 2016. Revisiting the data life cycle with big data curation. *International Journal of Digital Curation*, 10(2);180.

Press, G. 2014. 12 Big data definitions: What's yours? [online]. Available at: https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#171a 4abc13ae. [Accessed: 29 January 2018].

Pryor, G. (A) 2012. Big data – no big deal for curation? *Digital Curation Centre*, 1-33.

Pryor, G. (B) 2012. Managing research data. 7 Ridgmount Street, London: Facet Publishing.

Jones, S., Pryor, G., & Whyte, A. 2013. How to develop research data management services - a guide for HEIs. A digital curation centre 'working level' guide. *Digital Curation Centre.* 2.

Rahadi, P.S., Shobirin, K.A., & Ariyani, S. 2016. Big data management. *International Journal of Engineering and Emerging Technology* 1(1): 45-47.

Rauber, A., Asmi, A., Van Uytvanck, D., & Pröll, S. 2015. Data citation of evolving data. Recommendations of the Working Group on Data Citation (WGDC). Research Data Alliance [online] Available at: https://www.rd-alliance.org/system/files/RDA-DC-Recommendations_151020.pdf. [Accessed: 23 April 2018].

Research Data Alliance. 2018. RDA | Research data sharing without barriers. [online]. Available at: https://www.rd-alliance.org/. [Accessed: 16 August 2018].

Rossi, B. 2015. How to measure the value of big data. Information age. [online]. Available at: https://www.information-age.com/how-measure-value-big-data-123460041/. [Accessed: 7 October 2018].

Rossi, R., & Hirama, K. 2015. Characterising big data management. *Issues in Informing Science and Information Technology* 12: 165-180.

Rouse, M. 2013. Data stewardship. [ONLINE]. Available at: http://searchdatamanagement.techtarget.com/definition/data-stewardship. [Accessed: 29 January 2018].

Russom, P. 2013. Managing big data. TDWI Best Practices Report. Fourth Quarter 2013. *The Data Warehousing Institute*, 4-36.

Russom, P. 2017. Modern data integration and data quality practices for digital business requirements. *Transforming Data with Intelligence* 2-8.

Saagie. 2017. Data Governance, who's classifying your data? A data steward A data steward for your data classification. [online]. Available at: https://www.saagie.com/blog/a-data-steward-for-your-data-classification/. [Accessed: 26 November 2019].

Saltz, J.S. 2016. The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. *Big Data (Big Data) 2016 IEEE International Conference*, 2872-2879.

Schoonenboom, J., & Johnson, R.B. 2017. How to construct a mixed-methods research design. *Kolner Zeitschrift fur Soziolgie und Sozialpsychologie*. 69: 107-131.

Segment. 2018. Naming conventions: One step towards clean data. [online]. Available at: https://segment.com/academy/collecting-data/naming-conventions-for-clean-data/. [Accessed: 18 April 2018].

Seiner, R.S. 2013. Real-world data governance. What is a data steward and what do they do? *DataVersity*, 15-16.

Shafer, T. 2017. The 42 Vs of big data and data science. Elder research. [online]. Available at: https://www.elderresearch.com/blog/42-v-of-big-data. [Accessed: 7 October 2018].

Siljeur, A. 2018. DataFirst – Data curation process. [online]. Available at: https://www.datafirst.uct.ac.za/services/data-curation-process. [Accessed: 16 August 2018].

Singh, R. 2010. Case study method. [onlinE]. Available at: https://www.slideshare.net/SimplifyMyTraining/advantages-and-disadvantages-of-case-studies. [Accessed: 19 February 2018].

Solove, D. 2014. The Privacy Pillory and the Security Rack: The Enforcement Toolkit. [ONLINE]. Available at: https://teachprivacy.com/category/training-privacy-awareness/page/3/. [Accessed: 23 April 2019].

Study.com. 2019. Types of Different Degree Levels. [ONLINE] Available at: https://study.com/different_degrees.html. [Accessed: 11 November 2019].

TechPolicy.2014. What does big data mean for intellectual property protection? [online]. Available at: http://www.techpolicy.com/WhatDoesBigDataMean-IP-Protection.aspx. [Accessed: 09 October 2018].

Teets, M., & Goldner, M. 2013. Libraries' role in curating and exposing big data. *Future Internet.* 5: 429-438.

The Knowledge Academy. 2018. Big data and analytics | Big data and analytics training - South Africa. [online]. Available at: https://www.theknowledgeacademy.com/za/courses/big-data-and-analytics-training/. [Accessed 23 April 2018].

Times Higher Education. 2018. The World University Rankings 2018. [ONLINE]. Available at: https://www.timeshighereducation.com/world-university-rankings/2018/world-ranking#survey-answer. [Accessed: 6 November 2018].

Times Higher Education (THE). 2019. World University Rankings 2018. [ONLINE]. Available at: https://www.timeshighereducation.com/world-university-rankings/2018/world-ranking#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats. [Accessed 26 November 2019].

Tyhurst. J. 2018. The impact of big data on an intellectual property literacy training program. *King Abdullah University of Science and Technology Library*, 1-6.

University of Cape Town. 2018. Master's in data science | Statistical Sciences. [online]. Available at: http://www.stats.uct.ac.za/stats/study/postgrad/masters/data-science. [Accessed 23 April 2018].

The University of Nebraska - Lincoln. 2019. The Definition of Competencies and Their Application at NU | Human Resources | Nebraska. [ONLINE]. Available at: https://hr.unl.edu/compensation/nuvalues/corecompetencies.shtml/. [Accessed: 26 November 2019].

University of the Witwatersrand, Johannesburg. 2018. Big data analytics - Wits University. [online]. Available at: https://www.wits.ac.za/course-finder/postgraduate/science/big-data-analytics/#top-of-page. [Accessed 23 April 2018].

Van de Groenendaal, H.2016. SA needs more training in handling of big data. EE Publishers. [Online]. Available at: https://www.ee.co.za/article/sa-needs-training-handling-big-data.html. [Accessed: 16 August 2018].

Van Den Eynden, V. 2012. Looking after and managing your research data. UK Data Archive. City, University of Essex. 14-19.

Wamba, S.F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D., 2015. How 'big data' can make big impact: Findings from a systematic review and a longitudinal case. *International Journal of Production Economics.* 5-6.

Washington, E. 2018. Why Data Governance is Crucial for Big Data Environments. ReadITQuick. [ONLINE]. Available at: https://www.readitquik.com/articles/data/why-data-governance-is-crucial-for-big-data-environments/. [Accessed: 30 October 2018].

Merriam-Webster. 2019. Skill | Definition of Skill by Merriam-Webster. [ONLINE]. Available at: https://www.merriam-webster.com/dictionary/skill. [Accessed: 26 November 2019].

Wilkinson, M.D., et al. 2016. The FAIR guiding principles for scientific data management and stewardship. Scientific Data. 3(1): 1-9.

Williams, C. 2007. Research methods. *Journal of Business & Economic Research.* 5(3): 65-72.

World Data System. 2019. Emerging standards and best practices for data sharing, interoperability, use, and stewardship — World Data System: Trusted Data Services for Global Science. [ONLINE]. Available at: https://www.icsu-wds.org/events/wds-events/emerging-standards-and-best-practices-for-data-sharing-interoperability-use-and-stewardship. [Accessed: 16 August 2018].

Wyse, S.E. 2012. Advantages and disadvantages of surveys. SnapSurveys. [ONLINE]. Available at: https://www.snapsurveys.com/blog/advantages-disadvantages-surveys/. [Accessed: 19 October 2017].

Wyse, S.E. 2014. Advantages and disadvantages of open questions in course evaluations. SnapSurveys. [ONLINE]. Available at: https://www.snapsurveys.com/blog/advantages-disadvantages-open-questions-course-eval/. [Accessed: 8 October 2017].

Zhang, N., & Yuan, Q.J. 2016. An overview of data governance. ResearchGate, 2-7.

Zeppieri, G., & George, S.Z. 2017. Patient-defined desired outcome, success criteria, and expectation in outpatient physical therapy: a longitudinal assessment. *BioMed Central the Open Access Publisher*. 15:29.

Zwitter, A. 2014. Big data ethics. *Big Data and Society,* 1-6.

**Appendix 1: Data collection instrument**

**Section 1:**

| **Name of institution** | |
|---|---|
| Region | |

| # | **Detail collected** | **Data collected** |
|---|---|---|
| 1 | Web address | |
| 2 | Date accessed | |
| 3 | Field(s) of study | Instruction: (Summarise the detail & unpack in section 2 below. |
| 4 | Modules of data-related study included in curriculum | Instruction: (Summarise the detail & unpack in section 2 below. |
| 5 | Content (Topics) of Data management curriculum | Instruction: (Summarise the detail & unpack in section 2 below. |
| 6 | Electives | Yes / No |
| 7 | Skills being addressed | Instruction: (Summarise the detail & unpack in section 2 below. |
| 8 | Content (Topics) of Big data curriculum | Instruction: (Summarise the detail & unpack in section 3 below. |
| 9 | Electives | |
| 10 | Skills being addressed | |

| | | |
|---|---|---|
| 11 | Prescribed work | Yes / No |
| 12 | Learning outcomes | |
| 13 | Additional notes | |

**Section 2: Data module detail**

[Instruction: Add separate table for each module]

Module 1

| Description | Data collected | Additional notes |
|---|---|---|
| Degree in which module is presented | | |
| Module name: | | |
| Year of instruction/study | | |
| Credits | | |
| Module outcomes | | |
| Module topics | | |
| Prescribed work / documents | | |

Module 2

| Description | Data collected | Additional notes |
|---|---|---|

| | | |
|---|---|---|
| Degree in which module is presented | | |
| Module name: | | |
| Year of instruction/study | | |
| Credits | | |
| Module outcomes | | |
| Module topics | | |
| Prescribed work / documents | | |

Module 3

| Description | Data collected | Additional notes |
|---|---|---|
| Degree in which module is presented | | |
| Module name: | | |
| Year of instruction/study | | |
| Credits | | |
| Module outcomes | | |
| Module topics | | |
| Prescribed work / documents | | |

**Appendix 2: University web sites selected**

Please note that the appendix indicates the institution, the country in which the institution is located, as well as its ranking, web address(es) consulted, as well as the date it was accessed for the purpose of this study.

It is also important to note that the University of British Columbia is not in any of the regions mentioned as it is located in Canada. The only reason for this inclusion is the University of British Columbia's ranking within the library and information management schools (number 4), making it too strong of a resource to not be considered for this study.

| Name of Institution | Country | Rank | Web Address | Date Accessed |
|---|---|---|---|---|
| Australian National University | Australia | 49 | https://programsandcourses.anu.edu.au/course/STAT3017 | 08/11/2018 |
| Monash University | Australia | 84 | (1) http://www.monash.edu/pubs/2018handbooks/units/FIT5202.html<br>(2) http://www.monash.edu/pubs/2018handbooks/aos/data-science/ug-it-data-science.html | 14/11/2018 |
| University of Melbourne | Australia | 32 | https://handbook.unimelb.edu.au/2018/courses/mc-datasc | 08/11/2018 |
| University of Queensland | Australia | 69 | (1) https://future-students.uq.edu.au/study/program/Master-of-Data-Science-5659 | 14/11/2018 |

| Name of Institution | Country | Rank | Web Address | Date Accessed |
|---|---|---|---|---|
| | | | (2) https://my.uq.edu.au/programs-courses/program_list.html?acad_prog=5659&year=2019 | |
| University of Sydney | Australia | 59 | https://sydney.edu.au/courses/subject-areas/spec/big-data-in-business.html | 08/11/2018 |
| University of Copenhagen | Europe (Denmark) | 116 | https://copenhagensummeruniversity.ku.dk/en/courses/bigdata/ | 8/11/2018 |
| Ecole Polytechnique | Europe (France) | 108 | (1) https://www.polytechnique.edu/en/content/lx-launches-three-new-academic-programs-september<br>(2) https://gargantua.polytechnique.fr/siatel-web/linkto/mlCYYYT(fYS | 8/11/2018 |
| Technical University of Munich | Europe (Germany) | 44 | https://www.tum.de/en/studies/degree-programs/detail/data-engineering-and-analytics-master-of-science-msc/ | 8/11/2018 |
| University of Amsterdam | Europe (Netherlands) | 62 | (1) http://www.uva.nl/programmas/open-programmas-fnwi/hpc-big-data/hpc-big-data.html<br>(2) https://hpc.uva.nl/Roadmaps/article/122/Distributed-systems-and-BigData-(6-ECTS) | 8/11/2018 |
| University of Zurich | Europe (Switzerland) | 90 | https://www.oec.uzh.ch/en/research/excellence/methodology/big-data.html | 8/11/2018 |

| Name of Institution | Country | Rank | Web Address | Date Accessed |
|---|---|---|---|---|
| University of Cape Town | South Africa | 156 | http://www.stats.uct.ac.za/stats/study/postgrad/masters/data-science/ | 14/11/2018 |
| University of Pretoria | South Africa | 601 | https://www.up.ac.za/en/school-of-information-technology/article/2324622/mit-in-big-data-science-stream-c | 14/11/2018 |
| University of Witwatersrand | South Africa | 201 | https://www.wits.ac.za/course-finder/postgraduate/science/big-data-analytics/ | 14/11/2018 |
| Imperial College London | UK (England) | 9 | http://www.imperial.ac.uk/design-engineering/study/meng/modules/year-2/big-data/ | 8/11/2018 |
| University College London | UK (England) | 14 | https://www.ucl.ac.uk/silva/spp/teaching/undergraduate/ug-modules/courses/dsbda | 8/11/2018 |
| University of Manchester | UK (England) | 57 | https://www.manchester.ac.uk/study/undergraduate/courses/2019/06246/bsc-information-technology-management-for-business/course-details/BMAN10982#course-unit-details | 8/11/2018 |
| University of Oxford | UK (England) | 1 | https://www.bdi.ox.ac.uk/study-1 | 7/11/2018 |
| University of Sheffield | UK (England) | 106 | https://www.sheffield.ac.uk/dcs/postgraduate-taught/data-analytics | 04/03/2019 |

| Name of Institution | Country | Rank | Web Address | Date Accessed |
|---|---|---|---|---|
| University of Edinburgh | UK (Scotland) | 29 | https://www.epcc.ed.ac.uk/online-courses/courses/online-courses/courses/practical-introduction-data-science | 8/11/2018 |
| California Institute of Technology (Caltech) | United States | 5 | https://infosci.cornell.edu/masters/mps/curriculum/courses | 25/10/2018 |
| Harvard University | United States | 6 | (1) http://daslab.seas.harvard.edu/classes/cs265/<br>(2) http://daslab.seas.harvard.edu/classes/cs265/files/syllabus.pdf | 16/10/2018 |
| Indiana University of Bloomington | United States | 145 | https://bulletins.iu.edu/iub/soic/2018-2019/graduate/degree-programs /master-of-data-science/index.shtml | 12/03/2019 |
| Massachusetts Institute of Technology (MIT) | United States | 4 | https://mitxpro.mit.edu/courses/course-v1:MITxPRO+DSx+4T2018/about | 16/10/2018 |
| Rutgers – State University of New Jersey | United States | 176 | https://www.cs.rutgers.edu/courses/principles-of-information-and-data-management | 02/04/2019 |

| Name of Institution | Country | Rank | Web Address | Date Accessed |
|---|---|---|---|---|
| Stanford University | United States | 3 | (1) https://www.gsb.stanford.edu/exec-ed/programs/big-data-strategic-decisions/curriculum<br>(2) https://www.gsb.stanford.edu/sites/gsb/files/ee-sample-schedule-data-2018.pdf | 16/10/2018 |
| Syracuse University | United States | 251 | https://ischool.syr.edu/academics/gradutae/masters-degrees/ms-in-information-management/ | 25/03/2019 |
| University of Chicago | United States | 10 | https://www.chicagobooth.edu/executiveeducation/programs/marketing-and-sales/big-data-and-marketing-analytics#BoothTab2 | 25/10/2018 |
| University of Illinois – Urbana Champaign | United States | 71 | https://cs.illinois.edu/sites/default/files/docs/syllabi/CS598_IS531_DataCuration.pdf | 04/03/2019 |
| University of Maryland – College Park | United States | 82 | (1) https://www.cs.umd.edu/class/spring2016/cmsc724/<br>(2) https://www.cs.umd.edu/class/spring2018/cmsc498K-642/ | 03/04/2019 |
| University of Michigan – Ann Arbor | United States | 20 | (1) https://midas.umich/certifictae/<br>(2) https://michiganross.umich.edu/courses/big-data-management-tools-and-techniques-10164 | 25/03/2019 |

| Name of Institution | Country | Rank | Web Address | Date Accessed |
|---|---|---|---|---|
| University of North Carolina – Chapel Hill | United States | 56 | (1) https://ils.unc.edu/courses/2016_spring/inls626_001/Syllabus-626-001-Fall2016.pdf<br>(2) https://ils.unc.edu/courses/2014_fall/inls523_001/523_syll.htm | 04/03/2019 |
| University of Texas – Austin | United States | 39 | (1) https://www.cs.utexas.edu/courses/378-big-data-programming<br>(2) https://www.cs.utexas/courses/347-data-management | 02/04/2019 |
| University of British Columbia | Canada | 37 | https://slais.ubc.ca/programs/specialisations/data-services/ | 12/03/2019 |

**Appendix 3: Learning outcomes**

Please note, as stated in section 4.2.13 of this chapter, that if there is no module stated, the learning outcome then applies to the entire course. In retrospect, if there is no course stated and only a module, the learning outcome only applies to the module. In the case where both a course and a module are stated, the learning outcome then applies to the specific module form the course.

It is also important to note that all learning outcomes which are highlighted in the table are those which apply to the curator and the curator's activities. Many of the identified learning outcomes in this case were not curation based. The curation-based learning outcomes will be the main learning outcomes and will be used further within the research study. The remaining outcomes will, however, be used again in chapter 6.

| **Institution** | **Course** | **Module** | **Learning Outcomes** |
|---|---|---|---|
| **Library & Information Science Institutions** | | | **To be able to:** |
| University of Michigan – Ann Arbor | Graduate course in Data Science | Big Data Management | • Construct large data sets that source data from multiple sources<br>• Construct data visualisation<br>• *No data curation outcomes* |
| Rutgers – State of New Jersey | Undergraduate course | Principles of Information and Data Management | • Contribute to the field of Computer Science<br>• Gain practical knowledge and skills from the course research project |

| Institution | Course | Module | Learning Outcomes |
|---|---|---|---|
| | | | • Conduct business planning and programming projects<br>• *No data curation outcomes* |
| **Library & Information Management Institutions** | | | |
| University of North Carolina – Chapel Hill | Postgraduate course in Databases | Introduction to Big Data and NoSQL | • Gain knowledge of non-traditional and large-scale data applicators<br>• Gian knowledge of properties of social networking<br>• Gain knowledge of the fundamentals of NoSQL systems<br>• Gain experience with NoSQL systems and Hadoop<br>• Work and gain knowledge of database management systems<br>• Work and gain knowledge of database design principles |

| Institution | Course | Module | Learning Outcomes |
|---|---|---|---|
| | | | • Work and gain knowledge of database design tools<br>• *No data curation outcomes* |
| University of Illinois – Urbana Champaign | Graduate course in the foundations of data curation | Data Curation | • Gain experience with abstraction in data management<br>• Be able to identify relationships between key abstraction activities<br>• Understand hierarchies and strategies for data transformation and transcoding<br>• Perform data derivation<br>• Perform data preservation strategies<br>• Work with and construct dataset identifiers and citations<br>• Describe management of heterogeneity, including schema matching techniques<br>• Explain the role metadata plays in data management and identify a variety of metadata schemes |

| Institution | Course | Module | Learning Outcomes |
|---|---|---|---|
| | | | • Describe common data behaviours of managers, programmers, scientists, and other users<br>• Summarize the role institutions, agencies, policies, and laws play in data curation |
| **United States Institutions** | | | |
| Massachusetts Institute of Technology (MIT) | Postgraduate master's certification in Data, Systems and Society | Data Science and Big Data Analytics | • Be able to apply data science techniques to the organisation<br>• Learn how to avoid pitfalls in big data analytics<br>• Deploy machine learning algorithms to mine data<br>• Be able to interpret analytical models for effective decision making<br>• Convert datasets into models through predictive analytics<br>• Recognise challenges with big data algorithms<br>• Learn how to effectively represent data<br>• *No data curation outcomes* |

| Institution | Course | Module | Learning Outcomes |
|---|---|---|---|
| Stanford University | Senior level certification in big data, strategic decisions: analysis to action course | | • Be able to design methodologies to develop big data solutions<br>• Gain more knowledge on big data<br>• Gain knowledge and experience with machine learning<br>• Gain knowledge and experience with artificial Intelligence<br>• Make use of conceptual frameworks to recognise the potential of data<br>• Gain knowledge and experience with networking |
| Harvard University | Postgraduate course in Big Data Systems | CS265: Data Systems Research for the Big Data era | • Become familiar with industry trends in big data systems<br>• Understand the trade of designing and implementing big data systems<br>• Become more knowledgeable in decision making for big data scenarios<br>• Develop research skills |

| Institution | Course | Module | Learning Outcomes |
|---|---|---|---|
| | | | • Perform programming, debugging, and performance profiling<br>• *No data curation outcomes* |
| University of Chicago | Certificate in Marketing | Big Data and Big Data Analytics | • Be able to interpret data for strategic decision making<br>• Be able to use analytical frameworks for marketing strategies<br>• Use algorithmic tools for digital and non-digital marketing goals<br>• Recognise big data within a firms marketing strategy<br>• *No data curation outcomes* |
| **Australian Institutions** | | | |
| University of Melbourne | Graduate Master's course in Data Science | | • Understand advanced tools and methods in data science<br>• Understand machine learning |

| Institution | Course | Module | Learning Outcomes |
|---|---|---|---|
| | | | • Gain ethical awareness for the use of data and big data<br>• Evaluate data to understand research data in data science and data science disciplines<br>• Adapt to the evolution of data science |
| Monash University | Undergraduate course in Data Science | FIT5202 – Data Processing for Big Data | • Compare data streaming methods such as sampling, sketching and hashing<br>• Apply spatial data methods<br>• Apply large scale graphs, vectors and document processing methods<br>• Hadoop and Spark (curator would need to understand this topic)<br>• Evaluate the suitability of different distributed technologies for big data processing<br>• Explain cloud computing environment (curator would need to understand this topic) |
| Australian National University | Undergraduate course in Big Data Statistics | STAT3017 | • Explain how statistical features of big data impact traditional statistical methods and theory |

| Institution | Course | Module | Learning Outcomes |
|---|---|---|---|
| | | | • Discuss Random Matrix theory and its application in statistics on large scale<br>• Summarise the theory of sequential prediction and management of streaming data<br>• Demonstrate the use of computational tools to work with big and streaming data sets<br>• *No data curation outcomes* |
| **South African Institutions** | | | |
| University of Pretoria | Master's course in Big Data Science | Big Data Science | • Apply big data science in different domains<br>• Understand machine and statistical learning<br>• Hadoop, Python, and Spark (curator would need to understand this topic)<br>• Information ethics with big data science<br>• Gain knowledge on mathematical optimisation for big data science<br>• Be able to govern structured and unstructured data |

| Institution | Course | Module | Learning Outcomes |
|---|---|---|---|
| | | | • Understand architects available for processing big data<br>• Understand the different research methods for big data science<br>• Gain practical experience working with big data life cycle<br>• Gain research-based experience |
| University of the Witwatersrand | Honours course in Big Data Analytics | | • Gain knowledge on all facets of big data analytics<br>• Apply and gain knowledge on machine learning and optimisation of statistics<br>• Understand the introduction to big data analytics<br>• *No data curation outcomes* |
| **United Kingdom Institutions** | | | |
| Imperial College London | Undergraduate course in Data Analysis | Big Data | • Understand and be able to apply data visualisation techniques |

| Institution | Course | Module | Learning Outcomes |
|---|---|---|---|
| | | | • Understand and be able to apply statistical analysis<br>• Understand and be able to apply machine learning methods<br>• Interpret results of data analysis<br>• Solve problems with analytical techniques<br>• Understand and be able to apply Python, METLAB<br>• Gain practical presentation skills<br>• *No data curation outcomes* |
| University College London | Undergraduate course in Political Science | POLS3003 – Data Science and Big Data Analytics | • Understand data science<br>• Be able to apply data analysis |
| University of Edinburgh | Postgraduate course in Data Science | Practical Introduction to Data Science | • Apply data analytics techniques<br>• Understand data infrastructures used for data analytics<br>• Gain experience working with data analytics, data science, and big data for understanding<br>• Understand data management initiatives |

| Institution | Course | Module | Learning Outcomes |
|---|---|---|---|
| | | | • Gain knowledge on data management infrastructure<br>• Understand research data management plans and processes<br>• Understand and gain knowledge on Python |
| University of Manchester | Honours course in Information Technology Management for Business | Information Systems in Business and Introduction to Big Data and their manipulation | • Understand importance of data for managing organisations through information systems<br>• Demonstrate knowledge for different mechanisms for characteristics of data and big data<br>• Be able to perform data manipulation<br>• Create effective data storage schemas<br>• Apply knowledge to real-world situations by data gathering and processing of big data |
| **European Institutions** | | | |
| University of Amsterdam | Open programme High Performance | HPC and Big Data (6 ECTS | • Learn how to use high performance computing facilities |

| Institution | Course | Module | Learning Outcomes |
|---|---|---|---|
| | Computing and Big Data | | • Address big data issues by using distributed systems and big data<br>• Use super computers and high-performance computing clouds<br>• Understand and gain knowledge on programming and local remote visualisation techniques<br>• Understand and gain knowledge on data management<br>• Understand and gain knowledge on data intensive computing (curator would need knowledge on this topic)<br>• Understand high performance computing (curator would need knowledge on this topic)<br>• Understand distributed computing (curator would need knowledge on this topic)<br>• *No data curation outcomes* |

| Institution | Course | Module | Learning Outcomes |
|---|---|---|---|
| University of Copenhagen | Summer course in Big Data Analysis – Tools and Methods | | • Set up basic big data analysis<br>• Become acquainted with tools such as data cleaning, statistical methods for large datasets, data stream analysis, finding patterns and outliers in big data<br>• *No data curation outcomes* |

**Appendix 4: Content of data management and big data curriculum (Academic level)**

| | Diploma/Certificate | Undergraduate degree | Postgraduate degree | Honours degree | Master's degree | Doctoral degree |
|---|---|---|---|---|---|---|
| Library & Information Science institutions | × | × | × | | × | |
| Library & Information Management institutions | | × | × | | × | |
| United States institutions | × | | × | | × | |
| Australian institutions | × | × | × | | × | |
| South African institutions | | | | × | × | |
| European institutions | | × | × | | x | x |
| United Kingdom institutions | | × | × | x | | |

**Subsection 4.1: Library & Information Science institutions content of data management & big data curriculum (Core / Elective)**

| | Core module | Elective module |
|---|---|---|
| **Diploma/Certificate** | | X |
| **Undergraduate degree** | X | X |
| **Postgraduate degree** | X | |
| **Master's degree** | X | X |

**Subsection 4.2: Library & Information Management institutions content of data management & big data curriculum (Core / Elective)**

| | Core module | Elective module |
|---|---|---|
| **Undergraduate degree** | X | X |
| **Postgraduate degree** | X | |
| **Master's degree** | X | X |

**Subsection 4.3: United States institutions content of data management & big data curriculum (Core / Elective)**

| | Core module | Elective module |
|---|---|---|
| **Diploma/Certificate** | X | |
| **Postgraduate degree** | X | |
| **Master's degree** | X | X |

**Subsection 4.4: Australian institutions content of data management & big data curriculum (Core / Elective)**

| | Core module | Elective module |
|---|---|---|
| **Diploma/Certificate** | X | X |
| **Undergraduate degree** | X | |
| **Postgraduate degree** | X | X |
| **Master's degree** | X | X |

**Subsection 4.5: South African institutions content of data management & big data curriculum (Core / Elective)**

| | Core module | Elective module |
|---|---|---|
| **Honour's degree** | X | X |
| **Master's degree** | X | X |

**Subsection 4.6: European institutions content of data management & big data curriculum (Core / Elective)**

| | Core module | Elective module |
|---|---|---|
| **Undergraduate degree** | X | |
| **Postgraduate degree** | X | |
| **Master's degree** | X | X |
| **Doctoral degree** | X | |

**Subsection 4.7: United Kingdom institutions content of data management and big data curriculum (Core / Elective)**

| | Core module | Elective module |
|---|---|---|
| **Undergraduate degree** | X | |
| **Postgraduate degree** | X | |
| **Honour's degree** | X | |

**Appendix 5: Content of data management curriculum (Topics)**

From the below table, it is important to note that the institutions selected from the United States, Australia, South Africa, Europe, as well as the United Kingdom all exclude those institutions from the Library and Information Science schools. This ensured that there was no overlap in reporting on the selected institutions.

| Data topics | Library & Information Science / Management Schools (N=10) | United States institutions (Excluding LIS/LIM) (N=5) | Australian institutions (Excluding LIS/LIM) (N=5) | South African institutions (Excluding LIS/LIM) (N=3) | European institutions (Excluding LIS/LIM) (N=5) | UK institutions (Excluding LIS/LIM) (N=5) |
|---|---|---|---|---|---|---|
| Actionable data | | × | | | | |
| Basic programming | × | | | | | |
| Bioinformatics | | | × | | | |
| Computer Science | | | | | | × |
| Computer Security | x | | | | | |
| Database design | | | | × | | |
| Database learning | | × | | | | |
| Database management | × | | × | | | × |
| Database management strategies | × | | | | | |

| Data topics | Library & Information Science / Management Schools (N=10) | United States institutions (Excluding LIS/LIM) (N=5) | Australian institutions (Excluding LIS/LIM) (N=5) | South African institutions (Excluding LIS/LIM) (N=3) | European institutions (Excluding LIS/LIM) (N=5) | UK institutions (Excluding LIS/LIM) (N=5) |
|---|---|---|---|---|---|---|
| Database programming | | | × | | | × |
| Data analysis | | | | | | × |
| Data analytics | × | × | × | × | × | |
| Data engineering | | | | | | × |
| Data ethics | | × | | | | |
| Data mining | | | × | | | |
| Data modelling | × | | | | | |
| Data networking | | | | | | × |
| Data science | | | × | | × | |
| Data science principles | × | | | | | |
| Data storage | | | | | | × |
| Data structures | | | × | | | |
| Data visualisation | | | | × | | |
| Hardware fundamentals | | | | | | × |

| Data topics | Library & Information Science / Management Schools (N=10) | United States institutions (Excluding LIS/LIM) (N=5) | Australian institutions (Excluding LIS/LIM) (N=5) | South African institutions (Excluding LIS/LIM) (N=3) | European institutions (Excluding LIS/LIM) (N=5) | UK institutions (Excluding LIS/LIM) (N=5) |
|---|---|---|---|---|---|---|
| Information analytics | × | | | | | |
| Machine learning | x | | | × | | |
| Software engineering | | | × | | | |
| Systems engineering | | | | | | × |

**Appendix 6: Content of big data management curriculum (Topics)**

| Big data topics | Library & Information Science / Management Schools (N=10) | United States institutions (N=5) | Australian institutions (N=5) | South African institutions (N=3) | European institutions (N=5) | UK institutions (N=5) |
|---|---|---|---|---|---|---|
| Advanced database systems | | | × | | | |
| Advanced statistics | | | | | x | |
| Big data algorithms | × | × | | | | |
| Big data analysis | | | | | × | × |
| Big data ethics | | × | | | | x |
| Big data infrastructure | × | | | | | |
| Big data initiatives | | × | | | | |
| Big data management | | | | × | | |
| Big data patterns | | | | | x | |
| Big data map reducing | | | | | x | |
| Big data models | | × | | | | |
| Big data science | | | | × | | |
| Big data systems | × | | | | | |

| Big data topics | Library & Information Science / Management Schools (N=10) | United States institutions (N=5) | Australian institutions (N=5) | South African institutions (N=3) | European institutions (N=5) | UK institutions (N=5) |
|---|---|---|---|---|---|---|
| Big data visualisation | | × | | | | |
| Business analytics | | | | | x | |
| Cloud computing | | | × | | x | |
| Component analysis | | | × | | | |
| Databases | x | × | | | | |
| Database management | | | | | x | |
| Database usage | x | | | | | |
| Database systems | | | | | x | |
| Data analysis | | | | | | x |
| Data analytics | × | × | × | | | |
| Data cleaning | | | | | x | |
| Data curation | × | | | | x | |
| Data ethics | × | | | × | | x |
| Data exploration | | | | | | x |

| Big data topics | Library & Information Science / Management Schools (N=10) | United States institutions (N=5) | Australian institutions (N=5) | South African institutions (N=3) | European institutions (N=5) | UK institutions (N=5) |
|---|---|---|---|---|---|---|
| Data instruments and devices | | | | | x | |
| Data manipulation | | | | | | x |
| Data mapping | | × | | | | |
| Data mining | | | × | | × | x |
| Data modelling | × | × | | | | x |
| Data platforms | | × | | × | | |
| Data processing | | | | | | x |
| Data provenance | | × | | | | |
| Data science | | | × | | | x |
| Data skipping | | × | | | | |
| Data security | | × | | | | |
| Data set tools | x | | | | | |
| Data standards | x | | | | | |
| Data streaming systems | x | | | | | |

| Big data topics | Library & Information Science / Management Schools (N=10) | United States institutions (N=5) | Australian institutions (N=5) | South African institutions (N=3) | European institutions (N=5) | UK institutions (N=5) |
|---|---|---|---|---|---|---|
| Data structure | | × | | | | |
| Data systems | | × | | | x | |
| Data visualisation | | | × | | | x |
| Dimension reduction | | | × | | | |
| Hadoop | × | | | | | |
| Heterogeneity | | | × | | | |
| Hive | × | | | | | |
| Introduction to big data | × | | | | | |
| Machine learning | | × | × | × | × | × |
| Map-reducing | × | | | | | |
| Mathematical optimisation | | | | × | | |
| Matrix analysis | | | × | | | |
| Metadata | × | | | | | |
| Networking | | × | | | x | |

| Big data topics | Library & Information Science / Management Schools (N=10) | United States institutions (N=5) | Australian institutions (N=5) | South African institutions (N=3) | European institutions (N=5) | UK institutions (N=5) |
|---|---|---|---|---|---|---|
| Parallel programming models | | | | | x | |
| Practical artificial intelligence | | | | | x | |
| Predictive modelling | | | | | | x |
| Probability theory | | | × | | | |
| Programming languages | x | × | | | | |
| Quantitative market research | | | | | x | |
| Recommendation systems | | × | | | | |
| Rule-based learning | | | | | | x |
| Statistical learning | | | × | × | × | × |
| Structured databases | | | | | | x |
| SQL | × | | | | | |
| Unstructured data | | × | | | | |
| Web search & text analysis | | | × | | x | |

**Appendix 7: Curation learning outcomes and mapped to skills**

| **Curation learning outcomes – as identified by the researcher** | **Skills (listed in chapter 2) mapped to the learning outcomes** |
|---|---|
| 1. Gain:<br>a. Gain more knowledge on big data<br>b. Gain experience with abstractions in data management<br>c. Gain ethical awareness for the use of data and big data | Independent worker<br>Attention to detail<br>Data usage skills<br>Working with data and data formatting skills<br>Technical data skills<br>Value of data assets<br>Discipline specific requirements<br>Secondary data use<br>Metadata skills<br>Access, sharing and dissemination skills<br>Research ethics |
| 2. Understand:<br>a. Understand and gain knowledge on data management<br>3. Understand and gain knowledge on data intensive computing | Secondary data use<br>Independent worker<br>Attention to detail<br>Writing data management plans<br>Administrative skills<br>Metadata skills |

| **Curation learning outcomes – as identified by the researcher** | **Skills (listed in chapter 2) mapped to the learning outcomes** |
|---|---|
| a. Understand high performance computing<br>b. Understand distributed computing<br>c. Understand Hadoop, Spark and Python<br>d. Understand data science<br>e. Understand data management initiatives<br>f. Understand research data management plans and processes<br>g. Understand hierarchies and strategies for data transformation and transcoding<br>h. Understand advanced tools and methods in data science | Project management skills<br>Data usage skills<br>Organisation skills<br>Technical data skills<br>Community-based data skills<br>Research ethics<br>Developing policy and procedural documents |

| **Curation learning outcomes – as identified by the researcher** | **Skills (listed in chapter 2) mapped to the learning outcomes** |
|---|---|
| i. Understand information ethics with big data science | |
| 4. Demonstrate:<br>a. Demonstrate knowledge for different mechanisms for characteristics of data and big data | Secondary data use<br>Time management<br>Attention to detail<br>Independent worker<br>Metadata skills<br>Citation of data skills<br>Data usage skills<br>Working with data and data formatting skills<br>Organisation skills<br>Technical data skills |
| 5. Make use of:<br>a. Make use of conceptual framework to recognise the potential of data | Value of data assets<br>Secondary data use<br>Independent worker<br>Attention to detail<br>Administrative skills<br>Appraisal skills |

| **Curation learning outcomes – as identified by the researcher** | **Skills (listed in chapter 2) mapped to the learning outcomes** |
|---|---|
| | Data usage skills<br>Working with data and data formatting skills<br>Organisation skills<br>Technical data skills |
| 6. Explain:<br>a. Explain the cloud computing environment<br>b. Explain the role metadata plays in data management and identify a variety of metadata schemas | Technical data skills<br>Attention to detail<br>Independent worker<br>Administrative skills<br>Metadata skills<br>Data usage skills<br>Working with data and data formatting skills |
| 7. Be able to:<br>a. Be able to identify relationships between key abstraction activities<br>b. Be able to govern structured and unstructured data | Value of data assets<br>Safety and security<br>Secondary data use<br>Time management<br>Independent worker<br>Attention to detail<br>Metadata skills |

| **Curation learning outcomes – as identified by the researcher** | **Skills (listed in chapter 2) mapped to the learning outcomes** |
|---|---|
| | Citation of data skills<br>Archiving skills<br>Data usage skills<br>Working with data and data formatting skills<br>Rep<br>Organisation skills<br>Access, sharing and dissemination skills<br>Technical data skills |
| 8. Create:<br>a. Create effective data storage schemas | Data life cycle<br>Safety and security<br>Secondary data use<br>Time management<br>Independent worker<br>Attention to detail<br>Archiving skills<br>Metadata skills<br>Citation of data skills<br>Data usage skills |

| **Curation learning outcomes – as identified by the researcher** | **Skills (listed in chapter 2) mapped to the learning outcomes** |
|---|---|
| | Working with data and data formatting skills<br>Managing data storage<br>Repository skills<br>Access, sharing and dissemination skills<br>Technical data skills |
| 9. Perform:<br>a. Perform data derivation<br>b. Perform data preservation strategies | Data life cycle<br>Value of data assets<br>Secondary data use<br>Time management<br>Independent worker<br>Attention to detail<br>Appraisal skills<br>Archiving skills<br>Metadata skills<br>Data usage skills<br>Working with data and data formatting skills<br>Managing data storage<br>Repository skills |

| **Curation learning outcomes – as identified by the researcher** | **Skills (listed in chapter 2) mapped to the learning outcomes** |
|---|---|
| | Organisation skills<br>Access, sharing and dissemination skills<br>Technical data skills |
| 10. Work with:<br>a. Work with and construct identities and citations | Value of data assets<br>Discipline specific requirements<br>Secondary data use<br>Time management<br>Independent worker<br>Attention to detail<br>Metadata skills<br>Citation of data skills<br>Data usage skills<br>Technical data skills |
| 11. Describe:<br>a. Describe management of heterogeneity, including schema matching techniques | Funder requirements<br>Discipline specific requirements<br>Secondary data use<br>Time management<br>Independent worker |

| **Curation learning outcomes – as identified by the researcher** | **Skills (listed in chapter 2) mapped to the learning outcomes** |
|---|---|
| b. Describe common data behaviours of managers, programmers, scientists and other users | Attention to detail<br>Building relationships<br>Writing data management plans<br>Metadata skills<br>Citation of data skills<br>Data usage skills<br>Working with data and data formatting skills<br>Access, sharing and dissemination skills<br>Training skills<br>Technical data skills |
| 12. Summarise:<br>a. Summarise the role institutions, agencies, policies, and laws play in data curation | Funder requirements<br>Research ethics<br>Attention to detail<br>Building relationships<br>Community-based data skills<br>Developing budgets<br>Writing data management plans<br>Administrative skills |

| **Curation learning outcomes – as identified by the researcher** | **Skills (listed in chapter 2) mapped to the learning outcomes** |
|---|---|
| | Developing policy and procedural documents<br>Project management skills<br>Working with data and data formatting skills |
| 13. Evaluate:<br>a. Evaluate data to understand research data in data science and data science disciplines | Discipline specific requirements<br>Research ethics<br>Time management<br>Independent worker<br>Attention to detail Administrative skills<br>Data usage skills<br>Data research skills<br>Working with data and data management skills<br>Technical data skills |